Reward Maximization Through Discrete Active Inference

Lancelot Da Costa
l.da-costa@imperial.ac.uk
Department of Mathematics, Imperial College London, London SW7 2AZ, U.K.

Noor Sajid
noor.sajid.18@ucl.ac.uk
Thomas Parr
thomas.parr.12@ucl.ac.uk
Karl Friston
k.friston@ucl.ac.uk
Wellcome Centre for Human Neuroimaging, University College London, London, WC1N 3AR, U.K.

Ryan Smith
rsmith@laureateinstitute.org
Laureate Institute for Brain Research, Tulsa, OK 74136, U.S.A.

Active inference is a probabilistic framework for modeling the behavior of biological and artificial agents, which derives from the principle of minimizing free energy. In recent years, this framework has been applied successfully to a variety of situations where the goal was to maximize reward, often offering comparable and sometimes superior performance to alternative approaches. In this article, we clarify the connection between reward maximization and active inference by demonstrating how and when active inference agents execute actions that are optimal for maximizing reward. Precisely, we show the conditions under which active inference produces the optimal solution to the Bellman equation, a formulation that underlies several approaches to model-based reinforcement learning and control. On partially observed Markov decision processes, the standard active inference scheme can produce Bellman optimal actions for planning horizons of 1 but not beyond. In contrast, a recently developed recursive active inference scheme (sophisticated inference) can produce Bellman optimal actions on any finite temporal horizon. We append the analysis with a discussion of the broader relationship between active inference and reinforcement learning.

1 Introduction

1.1 Active Inference.

Active inference is a normative framework for modeling intelligent behavior in biological and artificial agents. It simulates

Neural Computation **35, 807–852** (2023)
https://doi.org/10.1162/neco_a_01574

behavior by numerically integrating equations of motion thought to describe the behavior of biological systems, a description based on the free energy principle (Barp et al., 2022; Friston et al., 2022). Active inference comprises a collection of algorithms for modeling perception, learning, and decision making in the context of both continuous and discrete state spaces (Barp et al., 2022; Da Costa et al., 2020; Friston et al., 2021, 2010; Friston, Parr, et al., 2017). Briefly, building active inference agents entails (1) equipping the agent with a (generative) model of the environment, (2) fitting the model to observations through approximate Bayesian inference by minimizing variational free energy (i.e., optimizing an evidence lower bound Beal, 2003; Bishop, 2006; Blei et al., 2017; Jordan et al., 1998) and (3) selecting actions that minimize expected free energy, a quantity that that can be decomposed into risk (i.e., the divergence between predicted and preferred paths) and ambiguity, leading to context-specific combinations of exploratory and exploitative behavior (Millidge, 2021; Schwartenbeck et al., 2019). This framework has been used to simulate and explain intelligent behavior in neuroscience (Adams et al., 2013; Parr, 2019; Parr et al., 2021; Sajid et al., 2022), psychology and psychiatry (Smith, Khalsa, et al., 2021; Smith, Kirlic, Stewart, Touthang, Kuplicki, Khalsa, et al., 2021; Smith, Kirlic, Stewart, Touthang, Kuplicki, McDermott, et al., 2021; Smith, Kuplicki, Feinstein, et al., 2020; Smith, Kuplicki, Teed, et al., 2020; Smith, Mayeli, et al., 2021; Smith, Schwartenbeck, Stewart, et al., 2020; Smith, Taylor, et al., 2022), machine learning (Çatal et al., 2020; Fountas et al., 2020; Mazzaglia et al., 2021; Millidge, 2020; Tschantz et al., 2019; Tschantz, Millidge, et al., 2020), and robotics (Çatal et al., 2021; Lanillos et al., 2020; Oliver et al., 2021; Pezzato et al., 2020; Pio-Lopez et al., 2016; Sancaktar et al., 2020; Schneider et al., 2022).

1.2 Reward Maximization through Active Inference? In contrast, the traditional approaches to simulating and explaining intelligent behavior—stochastic optimal control (Bellman, 1957; Bertsekas & Shreve, 1996) and reinforcement learning (RL; Barto & Sutton, 1992)—derive from the normative principle of executing actions to maximize reward scoring the utility afforded by each state of the world. This idea dates back to expected utility theory (Von Neumann & Morgenstern, 1944), an economic model of rational choice behavior, which also underwrites game theory (Von Neumann & Morgenstern, 1944) and decision theory (Berger, 1985; Dayan & Daw, 2008). Several empirical studies have shown that active inference can successfully perform tasks that involve collecting reward, often (but not always) showing comparative or superior performance to RL (Cullen et al., 2018; Marković et al., 2021; Mazzaglia et al., 2021; Millidge, 2020; Paul et al., 2021; Sajid, Ball, et al., 2021; Smith, Kirlic, Stewart, Touthang, Kuplicki, Khalsa, et al., 2021; Smith, Kirlic, Stewart, Touthang, Kuplicki, McDermott, et al., 2021; Smith, Schwartenbeck, Stewart, et al., 2020; Smith, Taylor, et al., 2022; van der Himst & Lanillos, 2020) and marked improvements when interacting with volatile environments (Marković et al., 2021; Sajid, Ball, et al., 2021). Given the prevalence and historical pedigree of reward maximization, we ask: *How and when do active inference agents execute actions that are optimal with respect to reward maximization?*

1.3 Organization of Paper.

In this article, we explain (and prove) how and when active inference agents exhibit (Bellman) optimal reward-maximizing behavior.

For this, we start by restricting ourselves to the simplest problem: maximizing reward on a finite horizon Markov decision process (MDP) with known transition probabilities—a sequential decision-making task with complete information. In this setting, we review the backward-induction algorithm from dynamic programming, which forms the workhorse of many optimal control and model-based RL algorithms. This algorithm furnishes a Bellman optimal state-action mapping, which means that it provides provably optimal decisions from the point of view of reward maximization (see section 2).

We then introduce active inference on finite horizon MDPs (see section 3)—a scheme consisting of perception as inference followed by planning as inference, which selects actions so that future states best align with preferred states.

In section 4, we show how and when active inference maximizes reward in MDPs. Specifically, when the preferred distribution is a (uniform mixture of) Dirac distribution(s) over reward-maximizing trajectories, selecting action sequences according to active inference maximizes reward (see section 4.1). Yet active inference agents, in their standard implementation, can select actions that maximize reward only when planning one step ahead (see section 4.2). It takes a recursive, sophisticated form of active inference to select actions that maximize reward—in the sense of a Bellman optimal state-action mapping—on any finite time-horizon (see section 4.3).

In section 5, we introduce active inference on partially observable Markov decision processes with known transition probabilities—a sequential decision-making task where states need to be inferred from observations—and explain how the results from the MDP setting generalize to this setting.

In section 6, we step back from the focus on reward maximization and briefly discuss decision making beyond reward maximization, learning unknown environments and reward functions, and outstanding challenges in scaling active inference. We append this with a broader discussion of the relationship between active inference and reinforcement learning in appendix A.

Our findings are summarized in section 7.

All of our analyses assume that the agent knows the environmental dynamics (i.e., transition probabilities) and reward function. In appendix A, we discuss how active inference agents can learn their world model and rewarding states when these are initially unknown—and the broader relationship between active inference and RL.

2 Reward Maximization on Finite Horizon MDPs

In this section, we consider the problem of reward maximization in Markov decision processes (MDPs) with known transition probabilities.

2.1 Basic Definitions.

MDPs are a class of models specifying environmental dynamics widely used in dynamic programming, model-based RL, and more broadly in engineering and artificial intelligence (Barto & Sutton, 1992; Stone, 2019). They are used to simulate sequential decision-making tasks with the objective of maximizing a reward or utility function. An MDP specifies environmental dynamics unfolding in discrete space and time under the actions pursued by an agent.

Definition 1 (Finite Horizon MDP). *A finite horizon MDP comprises the following collection of data:*

- $\mathbb{S}$*, a finite set of states.*
- $\mathbb{T} = \{0, \ldots, T\}$*, a finite set that stands for discrete time.* T *is the temporal horizon (a.k.a. planning horizon).*
- $\mathbb{A}$*, a finite set of actions.*
- $P(s_t = s' \mid s_{t-1} = s, a_{t-1} = a)$*, the probability that action* $a \in \mathbb{A}$ *in state* $s \in \mathbb{S}$ *at time* $t-1$ *will lead to state* $s' \in \mathbb{S}$ *at time* t*.* s_t *are random variables over* $\mathbb{S}$ *that correspond to the state being occupied at time* $t = 0, \ldots, T$*.*
- $P(s_0 = s)$*, the probability of being at state* $s \in \mathbb{S}$ *at the start of the trial.*
- $R(s)$*, the finite reward received by the agent when at state* $s \in \mathbb{S}$*.*

The dynamics afforded by a finite horizon MDP (see Figure 1) can be written globally as a probability distribution over state trajectories $s_{0:T} := (s_0, \ldots, s_T)$*, given a sequence of actions* $a_{0:T-1} := (a_0, \ldots, a_{T-1})$*, which factorizes as*

$$P(s_{0:T} \mid a_{0:T-1}) = P(s_0) \prod_{\tau=1}^{T} P(s_\tau \mid s_{\tau-1}, a_{\tau-1}).$$

Remark 1 (On the Definition of Reward). More generally, the reward function can be taken to be dependent on the previous action and previous state: $R_a\,(s' \mid s)$ is the reward received after transitioning from state s to state s' due to action a (Barto & Sutton, 1992; Stone, 2019). However, given an MDP with such a reward function, we can recover our simplified setting by defining a new MDP where the new states comprise the previous action, previous state, and current state in the original MDP. By inspection, the resulting reward function on the new MDP depends only on the current state (i.e., $R(s)$).

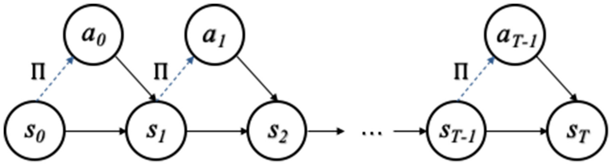

Figure 1: Finite horizon Markov decision process. This is a Markov decision process pictured as a Bayesian network (Jordan et al., 1998; Pearl, 1998). A finite horizon MDP comprises a finite sequence of states, indexed in time. The transition from one state to the next depends on action. As such, for any given action sequence, the dynamics of the MDP form a Markov chain on state-space. In this fully observed setting, actions can be selected under a state-action policy, Π, indicated with a dashed line: this is a probabilistic mapping from state-space and time to actions.

Remark 2 (Admissible Actions). In general, it is possible that only *some* actions can be taken at each state. In this case, one defines $\mathbb{A}_s$ to be the finite set of (allowable) actions from state $s \in \mathbb{S}$. All forthcoming results concerning MDPs can be extended to this setting.

To formalize what it means to choose actions in each state, we introduce the notion of a state-action policy.

Definition 2 (State-action Policy). *A state-action policy* Π *is a probability distribution over actions that depends on the state that the agent occupies, and time. Explicitly,*

$$\begin{aligned} \Pi : \mathbb{A} \times \mathbb{S} \times \mathbb{T} &\to [0,1] \\ (a, s, t) &\mapsto \Pi(a \mid s, t) \\ \forall (s,t) \in \mathbb{S} \times \mathbb{T} &: \sum_{a \in A} \Pi(a \mid s, t) = 1. \end{aligned}$$

When $s_t = s$, we will write $\Pi(a \mid s_t) := \Pi(a \mid s, t)$. Note that the action at the temporal horizon T is redundant, as no further can be reaped from the environment. Therefore, one often specifies state-action policies only up to time $T-1$, as $\Pi : \mathbb{A} \times \mathbb{S} \times \{0, \ldots, T-1\} \to [0,1]$. The state-action policy—as defined here—can be regarded as a generalization of a deterministic state-action policy that assigns the probability of 1 to an available action and 0 otherwise.

Remark 3 (Time-Dependent State-Action Policies). The way an agent chooses actions at the end of its life is usually going to be very different from the way it chooses them when it has a longer life ahead of it. In *finite* horizon decision problems, state-action policies should generally be considered to be time-dependent, as time-independent optimal state-action policies may not exist. To see this, consider the following simple example: $\mathbb{S} = \mathbb{Z}/5\mathbb{Z}$

(integers mod 5), $\mathbb{T} = \{0, 1, 2\}, \mathbb{A} = \{-1, 0, +1\}, R(0) = R(2) = R(3) = 0, R(1) = 1, R(4) = 6$. Optimal state-action policies are necessarily time-dependent as the reward-maximizing trajectory from state 2 at time 0 consists of reaching state 4, while the optimal trajectory from state 2 at time 1 consists of reaching state 1. This is particular to finite-horizon decisions, as, in infinite-horizon (discounted) problems, optimal state-action policies can always be taken to be time-independent (Puterman, 2014, theorem 6.2.7).

Remark 4 (Conflicting Terminologies: Policy in Active Inference). In active inference, a *policy* is defined as a sequence of actions indexed in time.[1] To avoid terminological confusion, we use action sequences to denote policies under active inference.

At time t, the goal is to select an action that maximizes future cumulative reward:

$$R(s_{t+1:T}) := \sum_{\tau=t+1}^{T} R(s_\tau).$$

Specifically, this entails following a state-action policy Π that maximizes the *state-value function*:

$$v_\Pi(s,t) := \mathbb{E}_\Pi[R(s_{t+1:T}) \mid s_t = s]$$

for any $(s,t) \in \mathbb{S} \times \mathbb{T}$. The state-value function scores the expected cumulative reward if the agent pursues state-action policy Π from the state $s_t = s$. When the state $s_t = s$ is clear from context, we will often write $v_\Pi(s_t) := v_\Pi(s,t)$. Loosely speaking, we will call the expected reward the *return*.

Remark 5 (Notation $\mathbb{E}_\Pi$). While standard in RL (Barto & Sutton, 1992; Stone, 2019), the notation $\mathbb{E}_\Pi[R(s_{t+1:T}) \mid s_t = s]$ can be confusing. It denotes the expected reward, under the transition probabilities of the MDP and a state-action policy Π, that is,

$$\mathbb{E}_{P(s_{t+1:T} \mid a_{t:T-1}, s_t = s)\Pi(a_{t:T-1} \mid s_{t+1:T-1}, s_t = s)}[R(s_{t+1:T})].$$

It is important to keep this correspondence in mind, as we will use both notations depending on context.

Remark 6 (Temporal Discounting). In infinite horizon MDPs (i.e., when T is infinite), RL often seeks to maximize the discounted sum of rewards,

[1] These are analogous to temporally extended actions or options introduced under the options framework in RL (Stolle & Precup, 2002).

$$v_\Pi(s,t) := \mathbb{E}_\Pi\left[\sum_{\tau=t}^{\infty}\gamma^{\tau-t}R(s_{\tau+1}) \mid s_t = s\right],$$

for a given temporal discounting term $\gamma \in (0,1)$ (Barto & Sutton, 1992; Bertsekas & Shreve, 1996; Kaelbling et al., 1998). In fact, temporal discounting is added to ensure that the infinite sum of future rewards converges to a finite value (Kaelbling et al., 1998). In finite horizon MDPs, temporal discounting is not necessary so we set $\gamma = 1$ (see Schmidhuber, 2006, 2010).

To find the best state-action policies, we would like to rank them in terms of their return. We introduce a partial ordering such that a state-action policy is *better* than another if it yields a higher return in any situation:

$$\Pi \geq \Pi' \iff \forall (s,t) \in \mathbb{S}\times\mathbb{T} : v_\Pi(s,t) \geq v_{\Pi'}(s,t).$$

Similarly, a state-action policy Π is strictly better than another Π' if it yields strictly higher returns:

$$\Pi > \Pi' \iff \Pi \geq \Pi' \text{ and } \exists (s,t) \in \mathbb{S}\times\mathbb{T} : v_\Pi(s,t) > v_{\Pi'}(s,t).$$

2.2 Bellman Optimal State-Action Policies. A state-action policy is Bellman optimal if it is better than all alternatives.

Definition 3 (Bellman Optimality). *A state-action policy Π^* is Bellman optimal if and only if it is better than all other state-action policies:*

$$\Pi^* \geq \Pi, \forall \Pi.$$

In other words, it maximizes the state-value function $v_\Pi(s,t)$ for any state s at time t.

It is important to verify that this concept is not vacuous.

Proposition 1 (Existence of Bellman Optimal State-Action Policies). *Given a finite horizon MDP as specified in definition 1, there exists a Bellman optimal state-action policy Π^*.*

A proof is found in appendix B.1. Note that the uniqueness of the Bellman optimal state-action policy is not implied by proposition 1; indeed, multiple Bellman optimal state-action policies may exist (Bertsekas & Shreve, 1996; Puterman, 2014).

Now that we know that Bellman optimal state-action policies exist, we can characterize them as a return-maximizing action followed by a Bellman optimal state-action policy.

Proposition 2 (Characterization of Bellman Optimal State-Action Policies). *For a state-action policy Π, the following are equivalent:*

1. *Π is Bellman optimal.*
2. *Π is both*
   a. *Bellman optimal when restricted to $\{1, \dots, T\}$. In other words, $\forall$ state-action policy Π' and $(s, t) \in \mathbb{S} \times \{1, \dots T\}$*

$$v_\Pi(s,t) \geq v_{\Pi'}(s,t).$$

   b. *At time 0, Π selects actions that maximize return:*

$$\Pi(a \mid s, 0) > 0 \iff a \in \arg\max_{a \in \mathbb{A}} \mathbb{E}_\Pi[R(s_{1:T}) \mid s_0 = s, a_0 = a], \quad \forall s \in \mathbb{S}. \tag{2.1}$$

A proof is in appendix B.2. Note that this characterization offers a recursive way to construct Bellman optimal state-action policies by successively selecting the best action, as specified by equation 2.1, starting from T and inducting backward (Puterman, 2014).

2.3 Backward Induction. Proposition 2 suggests a straightforward recursive algorithm to construct Bellman optimal state-action policies known as *backward induction* (Puterman, 2014). Backward induction has a long history. It was developed by the German mathematician Zermelo in 1913 to prove that chess has Bellman optimal strategies (Zermelo, 1913). In stochastic control, backward induction is one of the main methods for solving the Bellman equation (Adda & Cooper, 2003; Miranda & Fackler, 2002; Sargent, 2000). In game theory, the same method is used to compute subgame perfect equilibria in sequential games (Fudenberg & Tirole, 1991).

Backward induction entails planning backward in time, from a goal state at the end of a problem, by recursively determining the sequence of actions that enables reaching the goal. It proceeds by first considering the last time at which a decision might be made and choosing what to do in any situation at that time in order to get to the goal state. Using this information, one can then determine what to do at the second-to-last decision time. This process continues backward until one has determined the best action for every possible situation or state at every point in time.

Proposition 3 (Backward Induction: Construction of Bellman Optimal State-Action Policies). *Backward induction*

$$\begin{aligned}
\Pi(a \mid s, T-1) > 0 &\iff a \in \arg\max_{a \in \mathbb{A}} \mathbb{E}[R(s_T) \mid s_{T-1} = s, a_{T-1} = a], \quad \forall s \in \mathbb{S} \\
\Pi(a \mid s, T-2) > 0 &\iff a \in \arg\max_{a \in \mathbb{A}} \mathbb{E}_\Pi[R(s_{T-1:T}) \mid s_{T-2} = s, a_{T-2} = a], \\
&\qquad \forall s \in \mathbb{S} \\
&\qquad \vdots
\end{aligned}$$

$$\Pi(a \mid s, 0) > 0 \iff a \in \arg\max_{a \in \mathbb{A}} \mathbb{E}_{\Pi}[R(s_{1:T}) \mid s_0 = s, a_0 = a], \quad \forall s \in \mathbb{S} \tag{2.2}$$

defines a Bellman optimal state-action policy Π*. Furthermore, this characterization is complete: all Bellman optimal state-action policies satisfy the backward induction relation, equation 2.2.*

A proof is in appendix B.3.

Intuitively, the backward induction algorithm 2.2 consists of planning backward, by starting from the end goal and working out the actions needed to achieve the goal. To give a concrete example of this kind of planning, backward induction would consider the following actions in the order shown:

1. Desired goal: I would like to go to the grocery store.
2. Intermediate action: I need to drive to the store.
3. Current best action: I should put my shoes on.

Proposition 3 tells us that to be optimal with respect to reward maximization, one must plan like backward induction. This will be central to our analysis of reward maximization in active inference.

3 Active Inference on Finite Horizon MDPs

We now turn to introducing active inference agents on finite horizon MDPs with known transition probabilities. We assume that the agent's generative model of its environment is given by the previously defined finite horizon MDP (see definition 1). We do not consider the case where the transitions have to be learned but comment on it in appendix A.2 (see also Da Costa et al., 2020; Friston et al., 2016).

In what follows, we fix a time $t \geq 0$ and suppose that the agent has been in states $s_0, \dots, s_t$. To ease notation, we let $\vec{s} := s_{t+1:T}$, $\vec{a} := a_{t:T}$ be the future states and future actions. We define Q to be the *predictive distribution*, which encodes the predicted future states and actions given that the agent is in state s_t:

$$Q(\vec{s}, \vec{a} \mid s_t) := \prod_{\tau=t}^{T-1} Q(s_{\tau+1} \mid a_\tau, s_\tau) Q(a_\tau \mid s_\tau).$$

3.1 Perception as Inference.

In active inference, perception entails inferences about future, past, and current states given observations and a sequence of actions. When states are partially observed, this is done through variational Bayesian inference by minimizing a free energy functional also known as an evidence bound (Beal, 2003; Bishop, 2006; Blei et al., 2017; Wainwright & Jordan, 2007).

In the MDP setting, past and current states are known, so it is necessary only to infer future states given the current state and action sequence $P(\vec{s} \mid \vec{a}, s_t)$. These posterior distributions $P(\vec{s} \mid \vec{a}, s_t)$ can be computed exactly in virtue of the fact that the transition probabilities of the MDP are known; hence, variational inference becomes exact Bayesian inference:

$$Q(\vec{s} \mid \vec{a}, s_t) := P(\vec{s} \mid \vec{a}, s_t) = \prod_{\tau=t}^{T-1} P(s_{\tau+1} \mid s_\tau, a_\tau). \tag{3.1}$$

3.2 Planning as Inference. Now that the agent has inferred future states given alternative action sequences, we must assess these alternative plans by examining the resulting state trajectories. The objective that active inference agents optimize—in order to select the best possible actions—is the *expected free energy* (Barp et al., 2022; Da Costa et al., 2020; Friston et al., 2021). Under active inference, agents minimize expected free energy in order to maintain themselves distributed according to a target distribution C over the state-space $\mathbb{S}$ encoding the agent's preferences.

Definition 4 (Expected Free Energy on MDPs). *On MDPs, the expected free energy of an action sequence $\vec{a}$ starting from s_t is defined as (Barp et al., 2022, see section 5):*

$$G(\vec{a} \mid s_t) = \mathrm{D_{KL}}[Q(\vec{s} \mid \vec{a}, s_t) \mid C(\vec{s})], \tag{3.2}$$

where $\mathrm{D_{KL}}$ is the KL-divergence. Therefore, minimizing expected free energy corresponds to making the distribution over predicted states close to the distribution C that encodes prior preferences. Note that the expected free energy in partially observed MDPs comprises an additional ambiguity term (see section 5), which is dropped here as there is no ambiguity about observed states.

Since the expected free energy assesses the goodness of inferred future states under a course of action, we can refer to planning as inference (Attias, 2003; Botvinick & Toussaint, 2012). The expected free energy may be rewritten as

$$G(\vec{a} \mid s_t) = \underbrace{\mathbb{E}_{Q(\vec{s}\mid\vec{a},s_t)}[-\log C(\vec{s})]}_{\text{Expected surprise}} - \underbrace{\mathrm{H}[Q(\vec{s} \mid \vec{a}, s_t)]}_{\text{Entropy of future states}}. \tag{3.3}$$

Hence, minimizing expected free energy minimizes the expected surprise of states[2] according to C and maximizes the entropy of Bayesian beliefs over

[2] The surprise (also known as self-information or surprisal) of states—$\log C(\vec{s})$ is information-theoretic nomenclature (Stone, 2015) that scores the extent to which an

future states (a maximum entropy principle (Jaynes, 1957a), which is sometimes cast as keeping options open (Klyubin et al., 2008)).

Remark 7 (Numerical Tractability). The expected free energy is straightforward to compute using linear algebra. Given an action sequence $\vec{a}$, $C(\vec{s})$ and $Q(\vec{s} \mid \vec{a}, s_t)$ are categorical distributions over $\mathbb{S}^{T-t}$. Let their parameters be $\mathbf{c}, \mathbf{s}_{\vec{a}} \in [0,1]^{|\mathbb{S}|(T-1)}$, where $|\cdot|$ denotes the cardinality of a set. Then the expected free energy reads

$$G(\vec{a} \mid s_t) = \mathbf{s}_{\vec{a}}^{\mathrm{T}}(\log \mathbf{s}_{\vec{a}} - \log \mathbf{c}). \tag{3.4}$$

Notwithstanding, equation 3.4 is expensive to evaluate repeatedly when all possible action sequences are considered. In practice, one can adopt a temporal mean field approximation over future states (Millidge, Tschantz, & Buckley, 2020):

$$Q(\vec{s} \mid \vec{a}, s_t) = \prod_{\tau=t+1}^{T} Q(s_\tau \mid \vec{a}, s_{\tau-1}) \approx \prod_{\tau=t+1}^{T} Q(s_\tau \mid \vec{a}, s_t),$$

which yields the simplified expression

$$G(\vec{a} \mid s_t) \approx \sum_{\tau=t+1}^{T} \mathrm{D}_{\mathrm{KL}}[Q(s_\tau \mid \vec{a}, s_t) \mid C(s_\tau)]. \tag{3.5}$$

Expression 3.5 is much easier to handle: for each action sequence $\vec{a}$, one evaluates the summands sequentially $\tau = t+1, \ldots, T$, and if and when the sum up to τ becomes significantly higher than the lowest expected free energy encountered during planning, $G(\vec{a} \mid s_t)$ is set to an arbitrarily high value. Setting $G(\vec{a} \mid s_t)$ to a high value is equivalent to pruning away unlikely trajectories. This bears some similarity to decision tree pruning procedures used in RL (Huys et al., 2012). It finesses exploration of the decision tree in full depth and provides an Occam's window for selecting action sequences.

Complementary approaches can help make planning tractable. For example, hierarchical generative models factorize decisions into multiple levels. By abstracting information at a higher-level, lower levels entertain fewer actions (Friston et al., 2018), which reduces the depth of the decision tree by orders of magnitude. Another approach is to use algorithms that search the decision tree selectively, such as Monte Carlo tree search (Champion, Bowman, et al., 2021; Champion, Da Costa, et al., 2021; Fountas et al.,

observation is unusual under C. It does not imply that the agent experiences surprise in a subjective or declarative sense.

2020; Maisto et al., 2021; Silver et al., 2016) and amortizing planning using artificial neural networks (i.e., learning to plan) (Çatal et al., 2019; Fountas et al., 2020; Millidge, 2019; Sajid, Tigas, et al., 2021).

4 Reward Maximization on MDPs through Active Inference

Here, we show how active inference solves the reward maximization problem.

4.1 Reward Maximization as Reaching Preferences. From the definition of expected free energy, equation 3.2, active inference on MDPs can be thought of as reaching and remaining at a target distribution C over state-space.

The basic observation that underwrites the following is that the agent will maximize reward when the stationary distribution has all of its mass on reward maximizing states. To illustrate this, we define a preference distribution C_β, $\beta > 0$ over state-space $\mathbb{S}$, such that preferred states are rewarding states:[3]

$$\begin{aligned} C_\beta(\sigma) &:= \frac{\exp \beta R(\sigma)}{\sum_{\varsigma \in \mathbb{S}} \exp \beta R(\varsigma)} \propto \exp(\beta R(\sigma)), \quad \forall \sigma \in \mathbb{S} \\ \iff -\log C_\beta(\sigma) &= -\beta R(\sigma) - c(\beta), \\ &\qquad \forall \sigma \in \mathbb{S}, \text{ for some } c(\beta) \in \mathbb{R} \text{ constant w.r.t } \sigma. \end{aligned}$$

The (inverse temperature) parameter $\beta > 0$ scores how motivated the agent is to occupy reward-maximizing states. Note that states $s \in \mathbb{S}$ that maximize the reward $R(s)$ maximize $C_\beta(s)$ and minimize $-\log C_\beta(s)$ for any $\beta > 0$.

Using the additive property of the reward function, we can extend C_β to a probability distribution over trajectories $\vec{\sigma} := (\sigma_1, \ldots, \sigma_T) \in \mathbb{S}^T$. Specifically, C_β scores to what extent a trajectory is preferred over another trajectory:

$$\begin{aligned} C_\beta(\vec{\sigma}) &:= \frac{\exp \beta R(\vec{\sigma})}{\sum_{\vec{\varsigma} \in \mathbb{S}^T} \exp \beta R(\vec{\varsigma})} = \prod_{\tau=1}^{T} \frac{\exp \beta R(\sigma_\tau)}{\sum_{\varsigma \in \mathbb{S}} \exp \beta R(\varsigma)} = \prod_{\tau=1}^{T} C_\beta(\sigma_\tau), \quad \forall \vec{\sigma} \in \mathbb{S}^T \\ \iff -\log C_\beta(\vec{\sigma}) &= -\beta R(\vec{\sigma}) - c'(\beta) = -\sum_{\tau=1}^{T} \beta R(\sigma_\tau) - c'(\beta), \quad \forall \vec{\sigma} \in \mathbb{S}^T, \end{aligned} \tag{4.1}$$

where $c'(\beta) := c(\beta)T \in \mathbb{R}$ is constant with regard to $\vec{\sigma}$.

[3] Note the connection with statistical mechanics: β is an inverse temperature parameter, $-R$ is a potential function, and C_β is the corresponding Gibbs distribution (Pavliotis, 2014; Rahme & Adams, 2019).

When preferences are defined in this way, the preference distribution assigns exponentially more mass to states or trajectories that have a higher reward. Put simply, for trajectories $\vec{\sigma}, \vec{\varsigma} \in \mathbb{S}^T$ with reward $R(\vec{\sigma}) > R(\vec{\varsigma})$, the ratio of preference mass will be the exponential of the weighted difference in reward, where the weight is the inverse temperature:

$$\frac{C_\beta(\vec{\sigma})}{C_\beta(\vec{\varsigma})} = \frac{\exp(\beta R(\vec{\sigma}))}{\exp(\beta R(\vec{\varsigma}))} = \exp(\beta(R(\vec{\sigma}) - R(\vec{\varsigma}))). \tag{4.2}$$

As the temperature tends to zero, the ratio diverges so that $C_\beta(\vec{\sigma})$ becomes infinitely larger than $C_\beta(\vec{\varsigma})$. As C_β is a probability distribution (with a maximal value of one), we must have $C_\beta(\vec{\varsigma}) \overset{\beta\to+\infty}{\longrightarrow} 0$ for any suboptimal trajectory $\vec{\varsigma}$ and positive preference for reward maximizing trajectories (as all preferences must sum to one). In addition, all reward maximizing trajectories have the same probability mass by equation 4.2. Thus, in the zero temperature limit, preferences become a uniform mixture of Dirac distributions over reward-maximizing trajectories:

$$\lim_{\beta\to+\infty} C_\beta \propto \sum_{\vec{\sigma}\in I^{T-t}} \mathrm{Dirac}_{\vec{\sigma}}, \quad I := \arg\max_{s\in\mathbb{S}} R(s). \tag{4.3}$$

Of course, the above holds for preferences over individual states as it does for preferences over trajectories.

We now show how reaching preferred states can be formulated as reward maximization:

Lemma 1. *The sequence of actions that minimizes expected free energy also maximizes expected reward in the zero temperature limit* $\beta \to +\infty$ *(see equation 4.3):*

$$\lim_{\beta\to+\infty} \arg\min_{\vec{a}} G(\vec{a} \mid s_t) \subseteq \arg\max_{\vec{a}} \mathbb{E}_{Q(\vec{s}\mid\vec{a},s_t)}[R(\vec{s})].$$

Furthermore, of those action sequences that maximize expected reward, the expected free energy minimizers will be those that maximize the entropy of future states $\mathrm{H}[Q(\vec{s} \mid \vec{a}, s_t)]$.

A proof is in appendix B.4. In the zero temperature limit $\beta \to +\infty$, minimizing expected free energy corresponds to choosing the action sequence $\vec{a}$ such that $Q(\vec{s} \mid \vec{a}, s_t)$ has most mass on reward-maximizing states or trajectories (see Figure 2). Of those reward-maximizing candidates, the minimizer of expected free energy maximizes the entropy of future states $\mathrm{H}[Q(\vec{s} \mid \vec{a}, s_t)]$, thus keeping options open.

4.2 Reward Maximization on MDPs with a Temporal Horizon of 1. In this section, we first consider the case of a single-step decision problem (i.e., a temporal horizon of $T = 1$) and demonstrate how the standard active

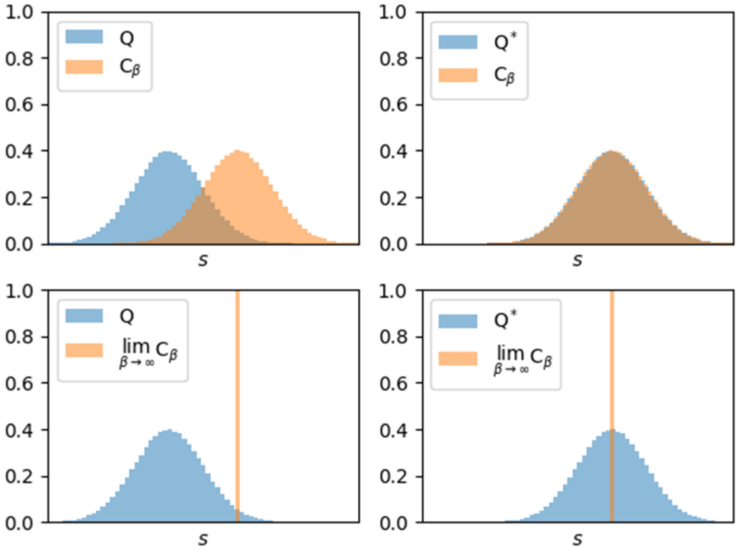

Figure 2: Reaching preferences and the zero temperature limit. We illustrate how active inference selects actions such that the predictive distribution $Q(\vec{s} \mid \vec{a}, s_t)$ most closely matches the preference distribution $C_\beta(\vec{s})$ (top right). We illustrate this with a temporal horizon of one, so that state sequences are states, which are easier to plot, but all holds analogously for sequences of arbitrary finite length. In this example, the state-space is a discretization of a real interval, and the predictive and preference distributions have a gaussian shape. The predictive distribution Q is assumed to have a fixed variance with respect to action sequences, such that the only parameter that can be optimized by action selection is its mean. In the zero temperature limit, equation 4.3, C_β becomes a Dirac distribution over the reward-maximizing state (bottom). Thus, minimizing expected free energy corresponds to selecting the action such that the predicted states assign most probability mass to the reward-maximizing state (bottom-right). Here, $Q^* := Q(\vec{s} \mid \vec{a}^*, s_t)$ denotes the predictive distribution over states given the action sequence that minimizes expected free energy $\vec{a}^* = \arg\min_{\vec{a}} G(\vec{a} \mid s_t)$.

inference scheme maximizes reward on this problem in the limit $\beta \to +\infty$. This will act as an important building block for when we subsequently consider more general multistep decision problems.

The standard decision-making procedure in active inference consists of assigning each action sequence with a probability given by the softmax of the negative expected free energy (Barp et al., 2022; Da Costa et al., 2020; Friston, FitzGerald, et al., 2017):

$$Q(\vec{a} \mid s_t) \propto \exp(-G(\vec{a} \mid s_t)).$$

Table 1: Standard Active Inference Scheme on Finite Horizon MDPs (Barp et al., 2022, section 5).

Process	Computation
Perceptual inference	$Q(\vec{s} \mid \vec{a}, s_t) = P(\vec{s} \mid \vec{a}, s_t) = \prod_{\tau=t}^{T-1} P(s_{\tau+1} \mid s_\tau, a_\tau)$
Planning as inference	$G(\vec{a} \mid s_t) = \mathrm{D_{KL}}[Q(\vec{s} \mid \vec{a}, s_t) \mid C(\vec{s})]$
Decision making	$Q(\vec{a} \mid s_t) \propto \exp(-G(\vec{a} \mid s_t))$
Action selection	$a_t \in \arg\max_{a \in \mathbb{A}} \left[Q(a_t = a \mid s_t) = \sum_{\vec{a}} Q(a_t = a \mid \vec{a}) Q(\vec{a} \mid s_t) \right]$

Agents then select the most likely action under this distribution:

$$a_t \in \arg\max_{a \in \mathbb{A}} Q(a \mid s_t) = \arg\max_{a \in \mathbb{A}} \sum_{\vec{a}} Q(a \mid \vec{a}) Q(\vec{a} \mid s_t)$$

$$= \arg\max_{a \in \mathbb{A}} \sum_{\vec{a}} Q(a \mid \vec{a}) \exp(-G(\vec{a} \mid s_t)) = \arg\max_{a \in \mathbb{A}} \sum_{\substack{\vec{a} \\ (\vec{a})_t = a}} \exp(-G(\vec{a} \mid s_t)).$$

In summary, this scheme selects the first action within action sequences that, on average, maximize their exponentiated negative expected free energies. As a corollary, if the first action is in a sequence with a very low expected free energy, this adds an exponentially large contribution to the selection of this particular action. We summarize this scheme in Table 1.

Theorem 1. *In MDPs with known transition probabilities and in the zero temperature limit $\beta \to +\infty$ (4.3), the scheme of Table 1,*

$$a_t \in \lim_{\beta \to +\infty} \arg\max_{a \in \mathbb{A}} \sum_{\substack{\vec{a} \\ (\vec{a})_t = a}} \exp(-G(\vec{a} \mid s_t)),$$

$$G(\vec{a} \mid s_t) = \mathrm{D_{KL}}[Q(\vec{s} \mid \vec{a}, s_t) \mid C_\beta(\vec{s})], \tag{4.4}$$

is Bellman optimal for the temporal horizon $T = 1$.

A proof is in appendix B.5. Importantly, *the standard active inference scheme, equation 4.4, falls short in terms of Bellman optimality on planning horizons greater than one*; this rests on the fact that it does not coincide with backward induction. Recall that backward induction offers a complete description of Bellman optimal state-action policies (see proposition 3). In contrast, active inference plans by adding weighted expected free energies of each possible future course of action. In other words, unlike backward induction, it considers future courses of action beyond the subset that will subsequently minimize expected free energy, given subsequently encountered states.

4.3 Reward Maximization on MDPs with Finite Temporal Horizons. To achieve Bellman optimality on finite temporal horizons, we turn to the expected free energy of an action given future actions that also minimize expected free energy. To do this, we can write the expected free energy recursively, as the immediate expected free energy, plus the expected free energy that one would obtain by subsequently selecting actions that minimize expected free energy (Friston et al., 2021). The resulting scheme consists of minimizing an expected free energy defined recursively, from the last time step to the current time step. In finite horizon MDPs, this reads

$$\begin{aligned} G(a_{T-1} \mid s_{T-1}) &= \mathrm{D}_{\mathrm{KL}}[Q(s_T \mid a_{T-1}, s_{T-1}) \mid C_\beta(s_T)] \\ G(a_\tau \mid s_\tau) &= \mathrm{D}_{\mathrm{KL}}[Q(s_{\tau+1} \mid a_\tau, s_\tau) \mid C_\beta(s_{\tau+1})] \\ &\qquad + \mathbb{E}_{Q(a_{\tau+1}, s_{\tau+1} \mid a_\tau, s_\tau)}[G(a_{\tau+1} \mid s_{\tau+1})], \quad \tau = t, \dots, T-2, \end{aligned}$$

where, at each time step, actions are chosen to minimize expected free energy:

$$Q(a_{\tau+1} \mid s_{\tau+1}) > 0 \iff a_{\tau+1} \in \arg\min_{a \in \mathbb{A}} G(a \mid s_{\tau+1}). \tag{4.5}$$

To make sense of this formulation, we unravel the recursion,

$$\begin{aligned} G(a_t \mid s_t) &= \mathrm{D}_{\mathrm{KL}}[Q(s_{t+1} \mid a_t, s_t) \mid C_\beta(s_{t+1})] + \mathbb{E}_{Q(a_{t+1}, s_{t+1} \mid a_t, s_t)}[G(a_{t+1} \mid s_{t+1})] \\ &= \mathrm{D}_{\mathrm{KL}}[Q(s_{t+1} \mid a_t, s_t) \mid C_\beta(s_{t+1})] \\ &\quad + \mathbb{E}_{Q(a_{t+1}, s_{t+1} \mid a_t, s_t)}\left[\mathrm{D}_{\mathrm{KL}}[Q(s_{t+2} \mid a_{t+1}, s_{t+1}) \mid C_\beta(s_{t+2})]\right] \\ &\quad + \mathbb{E}_{Q(a_{t+1:t+2}, s_{t+1:t+2} \mid a_t, s_t)}[G(a_{t+2} \mid s_{t+2})] \\ &= \dots = \mathbb{E}_{Q(\vec{a}, \vec{s} \mid a_t, s_t)} \sum_{\tau=t}^{T-1} \mathrm{D}_{\mathrm{KL}}[Q(s_{\tau+1} \mid a_\tau, s_\tau) \mid C_\beta(s_{\tau+1})] \\ &= \mathbb{E}_{Q(\vec{a}, \vec{s} \mid a_t, s_t)} \mathrm{D}_{\mathrm{KL}}[Q(\vec{s} \mid \vec{a}, s_t) \mid C_\beta(\vec{s})], \end{aligned} \tag{4.6}$$

which shows that this expression is exactly the expected free energy under action a_t, if one is to pursue future actions that minimize expected free energy, equation 4.5. We summarize this "sophisticated inference" scheme in Table 2.

The crucial improvement over the standard active inference scheme (see Table 1) is that planning is now performed based on subsequent counterfactual actions that minimize expected free energy as opposed to considering all future courses of action. Translating this into the language of state-action policies yields $\forall s \in \mathbb{S}$:

Table 2: Sophisticated active inference scheme on finite horizon MDPs (Friston et al., 2021).

Process	Computation
Perceptual inference	$Q(s_{\tau+1} \mid a_\tau, s_\tau) = P(s_{\tau+1} \mid a_\tau, s_\tau)$
Planning as inference	$G(a_\tau \mid s_\tau) = \mathrm{D}_{\mathrm{KL}}[Q(s_{\tau+1} \mid a_\tau, s_\tau) \mid C_\beta(s_{\tau+1})] \ldots$ $\ldots + \mathbb{E}_{Q(a_{\tau+1}, s_{\tau+1} \mid a_\tau, s_\tau)}[G(a_{\tau+1} \mid s_{\tau+1})]$
Decision making	$Q(a_\tau \mid s_\tau) > 0 \iff a_\tau \in \arg\min_{a \in \mathbb{A}} G(a \mid s_\tau)$
Action selection	$a_t \sim Q(a_t \mid s_t)$

$$\begin{aligned} a_{T-1}(s) &\in \arg\min_{a \in \mathbb{A}} G(a \mid s_{T-1} = s) \\ a_{T-2}(s) &\in \arg\min_{a \in \mathbb{A}} G(a \mid s_{T-2} = s) \\ &\vdots \\ a_1(s) &\in \arg\min_{a \in \mathbb{A}} G(a \mid s_1 = s) \\ a_0(s) &\in \arg\min_{a \in \mathbb{A}} G(a \mid s_0). \end{aligned} \tag{4.7}$$

Equation 4.7 is strikingly similar to the backward induction algorithm (proposition 3), and indeed we recover backward induction in the limit $\beta \to +\infty$.

Theorem 2 (Backward Induction as Active Inference). *In MDPs with known transition probabilities and in the zero temperature limit $\beta \to +\infty$, equation 4.3, the scheme of Table 2,*

$$\begin{aligned} Q(a_\tau \mid s_\tau) > 0 \quad &\iff a_t \in \lim_{\beta \to +\infty} \arg\min_{a \in \mathbb{A}} G(a \mid s_\tau) \\ G(a_\tau \mid s_\tau) \ &= \mathrm{D}_{\mathrm{KL}}[Q(s_{\tau+1} \mid a_\tau, s_\tau) \mid C_\beta(s_{\tau+1})] \\ &\quad + \mathbb{E}_{Q(a_{\tau+1}, s_{\tau+1} \mid a_\tau, s_\tau)}[G(a_{\tau+1} \mid s_{\tau+1})], \end{aligned} \tag{4.8}$$

is Bellman optimal on any finite temporal horizon as it coincides with the backward induction algorithm from proposition 3. Furthermore, if there are multiple actions that maximize future reward, those that are selected by active inference also maximize the entropy of future states $\mathrm{H}[Q(\vec{s} \mid \vec{a}, a, s_0)]$.

Note that maximizing the entropy of future states keeps the agent's options open (Klyubin et al., 2008) in the sense of committing the least to a specified sequence of states. A proof of theorem 2 is in appendix B.6.

5 Generalization to POMDPs

Partially observable Markov decision processes (POMDPs) generalize MDPs in that the agent observes a modality o_t, which carries incomplete information about the current state s_t, as opposed to the current state itself.

Definition 5 (Finite Horizon POMDP). *A finite horizon POMDP is an MDP (see definition 1) with the following additional data:*

- *$\mathbb{O}$ is a finite set of observations.*
- *$P(o_t = o \mid s_t = s)$ is the probability that the state $s \in \mathbb{S}$ at time t will lead to the observation $o \in \mathbb{O}$ at time t. o_t are random variables over $\mathbb{O}$ that correspond to the observation being sampled at time $t = 0, \dots, T$.*

5.1 Active Inference on Finite Horizon POMDPs.

We briefly introduce active inference agents on finite horizon POMDPs with known transition probabilities (for more details, see Da Costa et al., 2020; Parr et al., 2022; Smith, Friston, et al., 2022). We assume that the agent's generative model of its environment is given by POMDP (see definition 5).[4]

Let $\vec{s} := s_{0:T}, \vec{a} := a_{0:T-1}$ be all states and actions (past, present, and future), let $\tilde{o} := o_{0:t}$ be the observations available up to time t, and let $\vec{o} := o_{t+1:T}$ be the future observations. The agent has a predictive distribution over states given actions

$$Q(\vec{s} \mid \vec{a}, \tilde{o}) := \prod_{\tau=0}^{T-1} Q(s_{\tau+1} \mid a_\tau, s_\tau, \tilde{o}),$$

which is continuously updated following new observations.

5.1.1 Perception as Inference. In active inference, perception entails inferences about (past, present, and future) states given observations and a sequence of actions. When states are partially observed, the posterior distribution $P(\vec{s} \mid \vec{a}, \tilde{o})$ is intractable to compute directly. Thus, one approximates it by optimizing a variational free energy functional $F_{\vec{a}}$ (also known as an evidence bound; Beal, 2003; Bishop, 2006; Blei et al., 2017; Wainwright & Jordan, 2007) over a space of probability distributions $Q(\cdot \mid \vec{a}, \tilde{o})$ called the *variational family*:

$$P(\vec{s} \mid \vec{a}, \tilde{o}) = \arg\min_Q F_{\vec{a}}[Q(\vec{s} \mid \vec{a}, \tilde{o})] = \arg\min_Q \mathrm{D}_{\mathrm{KL}}[Q(\vec{s} \mid \vec{a}, \tilde{o}) \mid P(\vec{s} \mid \vec{a}, \tilde{o})]$$

$$F_{\vec{a}}[Q(\vec{s} \mid \vec{a}, \tilde{o})] := \mathbb{E}_{Q(\vec{s}\mid\vec{a},\tilde{o})}[\log Q(\vec{s} \mid \vec{a}, \tilde{o}) - \log P(\tilde{o}, \vec{s} \mid \vec{a})]. \tag{5.1}$$

[4] We do not consider the case where the model parameters have to be learned but comment on it in appendix A.2 (details in Da Costa et al., 2020; Friston et al., 2016).

Here, $P(\tilde{o}, \vec{s} \mid \vec{a})$ is the POMDP, which is supplied to the agent, and $P(\vec{s} \mid \vec{a}, \tilde{o})$. When the free energy minimum (see equation 5.1) is reached, the inference is exact:

$$Q(\vec{s} \mid \vec{a}, \tilde{o}) = P(\vec{s} \mid \vec{a}, \tilde{o}). \tag{5.2}$$

For numerical tractability, the variational family may be constrained to a parametric family of distributions, in which case equality is not guaranteed:

$$Q(\vec{s} \mid \vec{a}, \tilde{o}) \approx P(\vec{s} \mid \vec{a}, \tilde{o}). \tag{5.3}$$

5.1.2 Planning as Inference. The objective that active inference minimizes in order the select the best possible courses of action is the *expected free energy* (Barp et al., 2022; Da Costa et al., 2020; Friston et al., 2021). In POMDPs, the expected free energy reads (Barp et al., 2022, section 5)

$$G(\vec{a} \mid \tilde{o}) = \underbrace{\mathrm{D}_{\mathrm{KL}}[Q(\vec{s} \mid \vec{a}, \tilde{o}) \mid C_\beta(\vec{s})]}_{\text{Risk}} + \underbrace{\mathbb{E}_{Q(\vec{s}|\vec{a},\tilde{o})}\mathrm{H}[P(\vec{o} \mid \vec{s})]}_{\text{Ambiguity}}.$$

The expected free energy on POMDPs is the expected free energy on MDPs plus an extra term called *ambiguity*. This ambiguity term accommodates the uncertainty implicit in partially observed problems. The reason that this resulting functional is called expected free energy is because it comprises a relative entropy (risk) and expected energy (ambiguity). The expected free energy objective subsumes several decision-making objectives that predominate in statistics, machine learning, and psychology, which confers it with several useful properties when simulating behavior (see Figure 3 for details).

5.2 Maximizing Reward on POMDPs. Crucially, our reward maximization results translate to the POMDP case. To make this explicit, we rehearse lemma 1 in the context of POMDPs.

Proposition 4 (Reward Maximization on POMDPs). *In POMDPs with known transition probabilities, provided that the free energy minimum is reached (see equation 5.2), the sequence of actions that minimizes expected free energy also maximizes expected reward in the zero temperature limit $\beta \to +\infty$ (see equation 4.3):*

$$\lim_{\beta \to +\infty} \arg\min_{\vec{a}} G(\vec{a} \mid \tilde{o}) \subseteq \arg\max_{\vec{a}} \mathbb{E}_{Q(\vec{s}|\vec{a},\tilde{o})}[R(\vec{s})].$$

Furthermore, of those action sequences that maximize expected reward, the expected free energy minimizers will be those that maximize the entropy of future

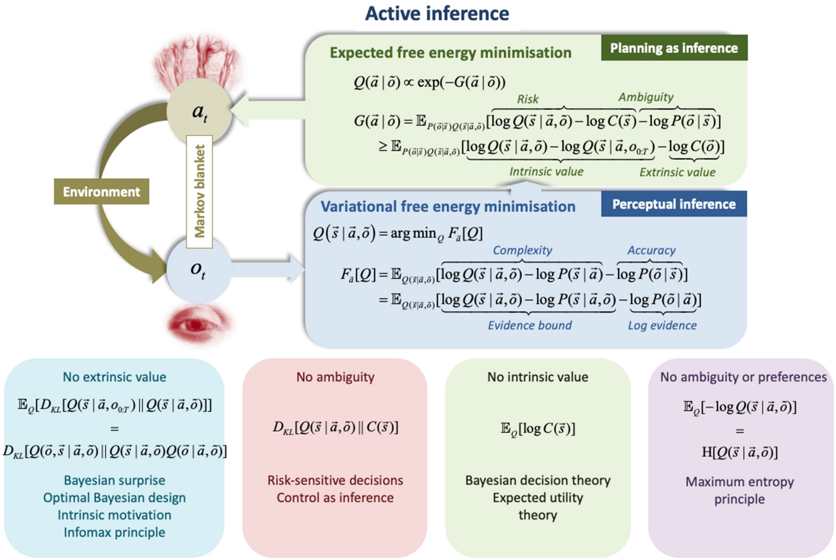
Active inference
Expected free energy minimisation
Planning as inference
$Q(\vec{a} \mid \tilde{o}) \propto \exp(-G(\vec{a} \mid \tilde{o}))$
$G(\vec{a} \mid \tilde{o}) = \mathbb{E}_{P(\tilde{o}|\vec{s})Q(\vec{s}|\vec{a},\tilde{o})}[\overbrace{\log Q(\vec{s} \mid \vec{a},\tilde{o}) - \log C(\vec{s})}^{Risk} \overbrace{- \log P(\tilde{o} \mid \vec{s})}^{Ambiguity}]$
$\geq \mathbb{E}_{P(\tilde{o}|\vec{s})Q(\vec{s}|\vec{a},\tilde{o})}[\underbrace{\log Q(\vec{s} \mid \vec{a},\tilde{o}) - \log Q(\vec{s} \mid \vec{a},o_{0:T})}_{Intrinsic\ value} - \underbrace{\log C(\tilde{o})}_{Extrinsic\ value}]$
Environment
a_t
Markov blanket
o_t
Variational free energy minimisation
Perceptual inference
$Q(\vec{s} \mid \vec{a},\tilde{o}) = \arg\min_Q F_{\vec{a}}[Q]$
$F_{\vec{a}}[Q] = \mathbb{E}_{Q(\vec{s}|\vec{a},\tilde{o})}[\overbrace{\log Q(\vec{s} \mid \vec{a},\tilde{o}) - \log P(\vec{s} \mid \vec{a})}^{Complexity} - \overbrace{\log P(\tilde{o} \mid \vec{s})}^{Accuracy}]$
$= \mathbb{E}_{Q(\vec{s}|\vec{a},\tilde{o})}[\underbrace{\log Q(\vec{s} \mid \vec{a},\tilde{o}) - \log P(\vec{s} \mid \vec{a},\tilde{o})}_{Evidence\ bound} - \underbrace{\log P(\tilde{o} \mid \vec{a})}_{Log\ evidence}]$
No extrinsic value
$\mathbb{E}_Q[D_{KL}[Q(\vec{s} \mid \vec{a},o_{0:T}) \| Q(\vec{s} \mid \vec{a},\tilde{o})]]$
=
$D_{KL}[Q(\tilde{o},\vec{s} \mid \vec{a},\tilde{o}) \| Q(\vec{s} \mid \vec{a},\tilde{o})Q(\tilde{o} \mid \vec{a},\tilde{o})]$
Bayesian surprise
Optimal Bayesian design
Intrinsic motivation
Infomax principle
No ambiguity
$D_{KL}[Q(\vec{s} \mid \vec{a},\tilde{o}) \| C(\vec{s})]$
Risk-sensitive decisions
Control as inference
No intrinsic value
$\mathbb{E}_Q[\log C(\vec{s})]$
Bayesian decision theory
Expected utility theory
No ambiguity or preferences
$\mathbb{E}_Q[-\log Q(\vec{s} \mid \vec{a},\tilde{o})]$
=
$\mathrm{H}[Q(\vec{s} \mid \vec{a},\tilde{o})]$
Maximum entropy principle

Figure 3: Active inference. The top panels illustrate the perception-action loop in active inference, in terms of minimization of variational and expected free energy. The lower panels illustrate how expected free energy relates to several descriptions of behavior that predominate in the psychological, machine learning, and economics. These descriptions are disclosed when one removes particular terms from the objective. For example, if we ignore extrinsic value, we are left with intrinsic value, variously known as expected information gain (Lindley, 1956; MacKay, 2003). This underwrites intrinsic motivation in machine learning and robotics (Barto et al., 2013; Deci & Ryan, 1985; Oudeyer & Kaplan, 2007) and expected Bayesian surprise in visual search (Itti & Baldi, 2009; Sun et al., 2011) and the organization of our visual apparatus (Barlow, 1961, 1974; Linsker, 1990; Optican & Richmond, 1987). In the absence of ambiguity, we are left with minimizing risk, which corresponds to aligning predicted states to preferred states. This leads to risk-averse decisions in behavioral economics (Kahneman & Tversky, 1979) and formulations of control as inference in engineering such as KL control (van den Broek et al., 2010). If we then remove intrinsic value, we are left with expected utility in economics (Von Neumann & Morgenstern, 1944) that underwrites RL and behavioral psychology (Barto & Sutton, 1992). Bayesian formulations of maximizing expected utility under uncertainty are also the basis of Bayesian decision theory (Berger, 1985). Finally, if we only consider a fully observed environment with no preferences, minimizing expected free energy corresponds to a maximum entropy principle over future states (Jaynes, 1957b, 1957a). Note that here $C(o)$ denotes the preferences over observations derived from the preferences over states. These are related by $P(o \mid s)C(s) = P(s \mid o)C(o)$.

states minus the (expected) entropy of outcomes given states $\mathrm{H}[Q(\vec{s} \mid \vec{a}, \tilde{o})] - \mathbb{E}_{Q(\vec{s}|a_t,\tilde{o})}\mathrm{H}[P(\vec{o} \mid \vec{s})]]$.

From proposition 4, we see that if there are multiple maximize reward action sequences, those that are selected maximize

$$\underbrace{\mathrm{H}[Q(\vec{s} \mid \vec{a}, \tilde{o})]}_{\text{Entropy of future states}} \quad - \quad \underbrace{\mathbb{E}_{Q(\vec{s}|a_t,\tilde{o})}[\mathrm{H}[P(\vec{o} \mid \vec{s})]]}_{\text{Entropy of observations given future states}} \quad .$$

In other words, they least commit to a prespecified sequence of future states and ensure that their expected observations are maximally informative of states. Of course, when inferences are inexact, the extent to which proposition 4 holds depends on the accuracy of the approximation, equation 5.3. A proof of proposition 4 is in appendix B.7.

The schemes of Tables 1 and 2 exist in the POMDP setting, (e.g., Barp et al., 2022, section 5, and Friston et al., 2021, respectively). Thus, in POMDPs with known transition probabilities, provided that inferences are exact (see equation 5.2) and in the zero temperature limit $\beta \to +\infty$ (see equation 4.3), standard active inference (Barp et al., 2022, section 5) maximizes reward on temporal horizons of one but not beyond, and a recursive scheme such as sophisticated active inference (Friston et al., 2021) maximizes reward on finite temporal horizons. Note that for computational tractability, the sophisticated active inference scheme presented in Friston et al. (2021) does not generally perform exact inference; thus, the extent to which it will maximize reward in practice will depend on the accuracy of its inferences. Nevertheless, our results indicate that sophisticated active inference will vastly outperform standard active inference in most reward-maximization tasks.

6 Discussion

In this article, we have examined a specific notion of optimality, namely, Bellman optimality, defined as selecting actions to maximize future expected rewards. We demonstrated how and when active inference is Bellman optimal on finite horizon POMDPs with known transition probabilities and reward function.

These results highlight important relationships among active inference, stochastic control, and RL, as well as conditions under which they would and would not be expected to behave similarly (e.g., environments with multiple reward-maximizing trajectories, those affording ambiguous observations). We refer readers to appendix A for a broader discussion of the relationship between active inference and reinforcement learning.

6.1 Decision Making beyond Reward Maximization. More broadly, it is important to ask if reward maximization is the right objective underwriting intelligent decision making. This is an important question for decision neuroscience. That is, do humans optimize a reward signal, expected free energy, or other planning objectives? This can be addressed by comparing the evidence for these competing hypotheses based on empirical data (Smith, Kirlic, Stewart, Touthang, Kuplicki, Khalsa, et al., 2021; Smith, Kirlic, Stewart, Touthang, Kuplicki, McDermott, et al., 2021; Smith, Schwartenbeck, Stewart, et al., 2020; Smith, Taylor, et al., 2022). Current empirical evidence suggests that humans are not purely reward-maximizing agents; they also engage in both random and directed exploration (Daw et al., 2006; Gershman, 2018; Mirza et al., 2018; Schulz & Gershman, 2019; Wilson et al., 2021, 2014; Xu et al., 2021) and keep their options open (Schwartenbeck, FitzGerald, Mathys, Dolan, Kronbichler, et al., 2015). As we have illustrated, active inference implements a clear form of directed exploration through minimizing expected free energy. Although not covered in detail here, active inference can also accommodate random exploration by sampling actions from the posterior belief over action sequences, as opposed to selecting the most likely action as presented in Tables 1 and 2.

Note that behavioral evidence favoring models that do not solely maximize reward within reward-maximization tasks—that is, where "maximize reward" is the explicit instruction—is not a contradiction. Rather, gathering information about the environment (exploration) generally helps to reap more reward in the long run, as opposed to greedily maximizing reward based on imperfect knowledge (Cullen et al., 2018; Sajid, Ball, et al., 2021). This observation is not new, and many approaches to simulating adaptive agents employed today differ significantly from their reward-maximizing antecedents (see appendix A.3).

6.2 Learning. When the transition probabilities or reward function are unknown to the agent, the problem becomes one of reinforcement learning (RL; Shoham et al., 2003 as opposed to stochastic control. Although we did not explicitly consider it above, this scenario can be accommodated by active inference by simply equipping the generative model with a prior and updating the model using variational Bayesian inference to best fit observed data. Depending on the specific learning problem and generative model structure, this can involve updating the transition probabilities and/or the target distribution C. In POMDPs it can also involve updating the probabilities of observations under each state. We refer to appendix A.2 for discussion of reward learning through active inference and connections to representative RL approaches, and Da Costa et al. (2020) and Friston et al. (2016) for learning transition probabilities through active inference.

6.3 Scaling Active Inference. When comparing RL and active inference approaches generally, one outstanding issue for active inference is whether it can be scaled up to solve the more complex problems currently handled by RL in machine learning contexts (Çatal et al., 2020, 2021; Fountas et al., 2020; Mazzaglia et al., 2021; Millidge, 2020; Tschantz et al., 2019). This is an area of active research.

One important issue along these lines is that planning ahead by evaluating all or many possible sequences of actions is computationally prohibitive in many applications. Three complementary solutions have emerged: (1) employing hierarchical generative models that factorize decisions into multiple levels and reduce the size of the decision tree by orders of magnitude (Çatal et al., 2021; Friston et al., 2018; Parr et al., 2021); (2) efficiently searching the decision tree using algorithms like Monte Carlo tree search (Champion, Bowman, et al., 2021; Champion, Da Costa, et al., 2021; Fountas et al., 2020; Maisto et al., 2021; Silver et al., 2016); and (3) amortizing planning using artificial neural networks (Çatal et al., 2019; Fountas et al., 2020; Millidge, 2019; Sajid, Tigas, et al., 2021).

Another issue rests on learning the generative model. Active inference may readily learn the parameters of a generative model; however, more work needs to be done on devising algorithms for learning the structure of generative models themselves (Friston, Lin, et al., 2017; Smith, Schwartenbeck, Parr, et al., 2020). This is an important research problem in generative modeling, called Bayesian model selection or structure learning (Gershman & Niv, 2010; Tervo et al., 2016).

Note that these issues are not unique to active inference. Model-based RL algorithms deal with the same combinatorial explosion when evaluating decision trees, which is one primary motivation for developing efficient model-free RL algorithms. However, other heuristics have also been developed for efficiently searching and pruning decision trees in model-based RL (Huys et al., 2012; Lally et al., 2017). Furthermore, model-based RL suffers the same limitation regarding learning generative model structure. Yet RL may have much to offer active inference in terms of efficient implementation and the identification of methods to scale to more complex applications (Fountas et al., 2020; Mazzaglia et al., 2021).

7 Conclusion

In summary, we have shown that under the specification that the active inference agent prefers maximizing reward, equation 4.3:

1. On finite horizon POMDPs with known transition probabilities, the objective optimized for action selection in active inference (i.e., expected free energy) produces reward-maximizing action sequences when state estimation is exact. When there are multiple reward-maximizing candidates, this selects those sequences that maximize the entropy of future states—thereby keeping options open—and that minimize the ambiguity of future observations so that they are maximally informative. More generally, the extent to which action

sequences will be reward maximizing will depend on the accuracy of state estimation.

2. The standard active inference scheme (e.g., Barp et al., 2022, section 5) produces Bellman optimal actions for planning horizons of one when state estimation is exact but not beyond.
3. A sophisticated active inference scheme (e.g., Friston et al., 2021) produces Bellman optimal actions on any finite planning horizon when state estimation is exact. Furthermore, this scheme generalizes the well-known backward induction algorithm from dynamic programming to partially observed environments. Note that for computational efficiency, the scheme presented in Friston et al. (2021), does not generally perform exact state estimation; thus, the extent to which it will maximize reward in practice will depend on the accuracy of its inferences. Nevertheless, it is clear from our results that sophisticated active inference will vastly outperform standard active inference in most reward-maximization tasks.

Note that for computational tractability, the sophisticated active inference scheme presented in Friston et al. (2021) does not generally perform exact inference; thus, the extent to which it will maximize reward in practice will depend on the accuracy of its inferences. Nevertheless, it is clear from these results that sophisticated active inference will vastly outperform standard active inference in most reward-maximization tasks.

In conclusion, the sophisticated active inference scheme should be the method of choice when applying active inference to optimally solve the reward-maximization problems considered here.

Appendix A: Active Inference and Reinforcement Learning

This article considers how active inference can solve the stochastic control problem. In this appendix, we discuss the broader relationship between active inference and RL.

Loosely speaking, RL is the field of methodologies and algorithms that learn reward-maximizing actions from data and seek to maximize reward in the long run. Because RL is a data-driven field, algorithms are selected based on how well they perform on benchmark problems. This has produced a plethora of diverse algorithms, many designed to solve specific problems, each with its own strengths and limitations. This makes RL difficult to characterize as a whole. Thankfully, many approaches to model-based RL and control can be traced back to approximating the optimal solution to the Bellman equation (Bellman & Dreyfus, 2015; Bertsekas & Shreve, 1996) (although this may become computationally intractable in high dimensions; Barto & Sutton, 1992). Our results showed how and when decisions under active inference and such RL approaches are similar.

This appendix discusses how active inference and RL relate and differ more generally. Their relationship has become increasingly important to understand, as a growing body of research has begun to (1) compare the performance of active inference and RL models in simulated environments (Cullen et al., 2018; Millidge, 2020; Sajid, Ball, et al., 2021), (2) apply active inference to model human behavior on reward learning tasks (Smith, Kirlic, Stewart, Touthang, Kuplicki, Khalsa, et al., 2021; Smith, Kirlic, Stewart, Touthang, Kuplicki, McDermott, et al., 2021; Smith, Schwartenbeck, Stewart, et al., 2020; Smith, Taylor, et al., 2022), and (3) consider the complementary predictions and interpretations each offers in computational neuroscience, psychology, and psychiatry (Cullen et al., 2018; Huys et al., 2012; Schwartenbeck, FitzGerald, Mathys, Dolan, & Friston, 2015; Schwartenbeck et al., 2019; Tschantz, Seth, et al., 2020).

A.1 Main Differences between Active Inference and Reinforcement Learning.

A.1.1 Philosophy. Active inference and RL differ profoundly in their philosophy. RL derives from the normative principle of maximizing reward (Barto & Sutton, 1992), while active inference describes systems that maintain their structural integrity over time (Barp et al., 2022; Friston et al., 2022). Despite this difference, these frameworks have many practical similarities. For example, recall that behavior in active inference is completely determined by the agent's preferences, determined as priors in their generative model. Crucially, log priors can be interpreted as reward functions and vice versa, which is how behavior under RL and active inference can be related.

A.1.2 Model Based and Model Free. Active inference agents always embody a generative (i.e., forward) model of their environment, while RL comprises both model-based and simpler model-free algorithms. In brief, "model-free" means that agents learn a reward-maximizing state-action mapping, based on updating cached state-action pair values through initially random actions that do not consider future state transitions. In contrast, model-based RL algorithms attempt to extend stochastic control approaches by learning the dynamics and reward function from data. Recall that stochastic control calls on strategies that evaluate different actions on a carefully handcrafted forward model of dynamics (i.e., known transition probabilities) to finally execute the reward-maximizing action. Under this terminology, all active inference agents are model-based.

A.1.3 Modeling Exploration. Exploratory behavior—which can improve reward maximization in the long run—is implemented differently in the two approaches. In most cases, RL implements a simple form of exploration by incorporating randomness in decision making (Tokic & Palm, 2011; Wilson et al., 2014), where the level of randomness may or may not change over time as a function of uncertainty. In other cases, RL incorporates ad hoc information bonuses in the reward function or other decision-making objectives to build in directed exploratory drives (e.g., upper-confidence-bound algorithms or Thompson sampling). In contrast, directed exploration emerges naturally within active inference through interactions between the risk and ambiguity terms in the expected free energy (Da Costa et al., 2020; Schwartenbeck et al., 2019). This addresses the explore-exploit dilemma and confers the agent with artificial curiosity (Friston, Lin, et al., 2017; Schmidhuber, 2010; Schwartenbeck et al., 2019; Still & Precup, 2012), as opposed to the need to add ad hoc information bonus terms (Tokic & Palm, 2011). We expand on this relationship in appendix A.3.

A.1.4 Control and Learning as Inference. Active inference integrates state estimation, learning, decision making, and motor control under the single objective of minimizing free energy (Da Costa et al., 2020). In fact, active inference extends previous work on the duality between inference and control (Kappen et al., 2012; Rawlik et al., 2013; Todorov, 2008; Toussaint, 2009) to solve motor control problems via approximate inference (i.e., planning as inference: Attias, 2003; Botvinick & Toussaint, 2012; Friston et al., 2012, 2009; Millidge, Tschantz, Seth, et al., 2020). Therefore, some of the closest RL methods to active inference are control as inference, also known as maximum entropy RL (Levine, 2018; Millidge, Tschantz, Seth, et al., 2020; Ziebart, 2010), though one major difference is in the choice of decision-making objective. Loosely speaking, these aforementioned methods minimize the risk term of the expected free energy, while active inference also minimizes ambiguity.

Useful Features of Active Inference

1. Active inference allows great flexibility and transparency when modeling behavior. It affords explainable decision making as a mixture of information- and reward-seeking policies that are explicitly encoded (and evaluated in terms of expected free energy) in the generative model as priors, which are specified by the user (Da Costa, Lanillos, et al., 2022). As we have seen, the kind of behavior that can be produced includes the optimal solution to the Bellman equation.
2. Active inference accommodates deep hierarchical generative models combining both discrete and continuous state-spaces (Friston, Parr, et al., 2017; Friston et al., 2018; Parr et al., 2021).
3. The expected free energy objective optimized during planning subsumes many approaches used to describe and simulate decision making in the physical, engineering, and life sciences, affording it various interesting properties as an objective (see Figure 3 and Friston et al., 2021). For example, exploratory and exploitative behavior are canonically integrated, which finesses the need for manually incorporating

ad hoc exploration bonuses in the reward function (Da Costa, Tenka, et al., 2022).

4. Active inference goes beyond state-action policies that predominate in traditional RL to sequential policy optimization. In sequential policy optimization, one relaxes the assumption that the same action is optimal given a particular state and acknowledges that the sequential order of actions may matter. This is similar to the linearly solvable MDP formulation presented by Todorov (2006, 2009), where transition probabilities directly determine actions and an optimal policy specifies transitions that minimize some divergence cost. This way of approaching policies is perhaps most apparent in terms of exploration. Put simply, it is clearly better to explore and then exploit than the converse. Because expected free energy is a functional of beliefs, exploration becomes an integral part of decision making—in contrast with traditional RL approaches that try to optimize a reward function of states. In other words, active inference agents will explore until enough uncertainty is resolved for reward-maximizing, goal-seeking imperatives to start to predominate.

Such advantages should motivate future research to better characterize the environments in which these properties offer useful advantages—such as where performance benefits from learning and planning at multiple temporal scales and from the ability to select policies that resolve both state and parameter uncertainty.

A.2 Reward Learning. Given the focus on relating active inference to the objective of maximizing reward, it is worth briefly illustrating how active inference can learn the reward function from data and its potential connections to representative RL approaches. One common approach for active inference to learn a reward function (Smith, Schwartenbeck, Stewart, et al., 2020; Smith, Taylor, et al., 2022) is to set preferences over observations rather than states, which corresponds to assuming that inferences over states given outcomes are accurate:

$$\begin{aligned}
\underbrace{\mathrm{D}_{\mathrm{KL}}\left[Q\left(\vec{s} \mid \vec{a}, \tilde{o}\right) \mid C\left(\vec{s}\right)\right]}_{\text{Risk (states)}} &= \underbrace{\mathrm{D}_{\mathrm{KL}}\left[Q\left(\vec{o} \mid \vec{a}, \tilde{o}\right) \mid C\left(\vec{o}\right)\right]}_{\text{Risk (outcomes)}} \\
&\quad + \underbrace{\mathbb{E}_{Q(\vec{o}\mid\vec{a},\tilde{o})}\left[\mathrm{D}_{\mathrm{KL}}\left[Q\left(\vec{s} \mid \vec{o}, \tilde{o}, \vec{a}\right) \mid P\left(\vec{s} \mid \vec{o}\right)\right]\right]}_{\approx 0} \\
&\approx \underbrace{\mathrm{D}_{\mathrm{KL}}\left[Q\left(\vec{o} \mid \vec{a}, \tilde{o}\right) \mid C\left(\vec{o}\right)\right]}_{\text{Risk (outcomes)}},
\end{aligned}$$

that is, equality holds whenever the free energy minimum is reached (see equation 5.2). Then one sets the preference distribution such that the observations designated as rewards are most preferred. In the zero temperature limit (see eqnarray 4.3), preferences only assign mass to reward-maximizing observations. When formulated in this way, the reward signal is treated as sensory data, as opposed to a separate signal from the environment. When one sets allowable actions (controllable state transitions) to be fully deterministic such that the selection of each action will transition the agent to a given state with certainty, the emerging dynamics are such that the agent chooses actions to resolve uncertainty about the probability of observing reward under each state. Thus, learning the reward probabilities of available actions amounts to learning the likelihood matrix $P(\vec{o} \mid \vec{s}) := o_t \cdot As_t$, where A is a stochastic matrix. This is done by setting a prior $\mathbf{a}$ over A, that is, a matrix of nonnegative components, the columns of which are Dirichlet priors over the columns of A. The agent then learns by accumulating Dirichlet parameters. Explicitly, at the end of a trial or episode, one sets (Da Costa et al., 2020; Friston et al., 2016)

$$\mathbf{a} \leftarrow \mathbf{a} + \sum_{\tau=0}^{T} o_\tau \otimes Q(s_\tau \mid o_{0:T}). \tag{A.1}$$

In equation A.1, $Q(s_\tau \mid o_{0:T})$ is seen as a vector of probabilities over the state-space $\mathbb{S}$, corresponding to the probability of having been in one or another state at time τ after having gathered observations throughout the trial. This rule simply amounts to counting observed state-outcome pairs, which is equivalent to state-reward pairs when the observation modalities correspond to reward.

One should not conflate this approach with the update rule consisting of accumulating state-observation counts in the likelihood matrix,

$$A \leftarrow A + \sum_{\tau=0}^{T} o_\tau \otimes Q(s_\tau \mid o_{0:T}), \tag{A.2}$$

and then normalizing its columns to sum to one when computing probabilities. The latter simply approximates the likelihood matrix A by accumulating the number of observed state-outcome pairs. This is distinct from the approach outlined above, which encodes uncertainty over the matrix A, as a probability distribution over possible distributions $P(o_t \mid s_t)$. The agent is initially very unconfident about A, which means that it doesn't place high-probability mass on any specification of $P(o_t \mid s_t)$. This uncertainty is gradually resolved by observing state-observation (or state-reward) pairs. Computationally, it is a general fact of Dirichlet priors that an increase in elements of $\mathbf{a}$ causes the entropy of $P(o_t \mid s_t)$ to decrease. As the terms added in equation A.1 are always positive, one choice of distribution

$P(o_t \mid s_t)$—which best matches available data and prior beliefs—is ultimately singled out. In other words, the likelihood mapping is learned.

The update rule consisting of accumulating state-observation counts in the likelihood matrix (see equation A.2) (i.e., not incorporating Dirichlet priors) bears some similarity to off-policy learning algorithms such as Q-learning. In Q-learning, the objective is to find the best action given the current observed state. For this, the Q-learning agent accumulates values for state-action pairs with repeated observation of rewarding or punishing action outcomes—much like state-observation counts. This allows it to learn the Q-value function that defines a reward maximizing policy.

As always in partially observed environments, we cannot guarantee that the true likelihood mapping will be learned in practice. Smith et al. (2019) provides examples where, although not in an explicit reward-learning context, learning the likelihood can be more or less successful in different situations. Learning the true likelihood fails when the inference over states is inaccurate, such as when using too severe a mean-field approximation to the free energy (Blei et al., 2017; Parr et al., 2019; Tanaka, 1999), which causes the agent to misinfer states and thereby accumulate Dirichlet parameters in the wrong locations. Intuitively, this amounts to jumping to conclusions too quickly.

Remark 8. If so desired, reward learning in active inference can also be equivalently formulated as learning transition probabilities $P(s_{t+1} \mid s_t, a_t)$. In this alternative setup (as exemplified in Sales et al. (2019)), mappings between reward states and reward outcomes in A are set as identity matrices, and the agent instead learns the probability of transitioning to states that deterministically generate preferred (rewarding) observations given the choice of each action sequence. The transition probabilities under each action are learned in a similar fashion as above (see equation A.1), by accumulating counts on a Dirichlet prior over $P(s_{t+1} \mid s_t, a_t)$. See Da Costa et al., 2020, appendix, for details.

Given the model-based Bayesian formulation of active inference, more direct links can be made between the active inference approach to reward learning described above and other Bayesian model-based RL approaches. For such links to be realized, the Bayesian RL agent would be required to have a prior over a prior (e.g., a prior over the reward function prior or transition function prior). One way to implicitly incorporate this is through Thompson sampling (Ghavamzadeh et al., 2016; Russo & Van Roy, 2014, 2016; Russo et al., 2017). While that is not the focus of this article, future work could further examine the links between reward learning in active inference and model-based Bayesian RL schemes.

A.3 Solving the Exploration-Exploitation Dilemma.

An important distinction between active inference and reinforcement learning schemes is how they solve the exploration-exploitation dilemma.

The exploration-exploitation dilemma (Berger-Tal et al., 2014) arises whenever an agent has incomplete information about its environment, such as when the environment is partially observed or the generative model has to be learned. The dilemma is then about deciding whether to execute actions aiming to collect reward based on imperfect information about the environment or to execute actions aiming to gather more information—allowing the agent to reap more reward in the future. Intuitively, it is always best to explore and then exploit, but optimizing this trade-off can be difficult.

Active inference balances exploration and exploitation through minimizing the risk and ambiguity inherent in the minimization of expected free energy. This balance is context sensitive and can be adjusted by modifying the agent's preferences (Da Costa, Lanillos, et al., 2022). In turn, the expected free energy is obtained from a description of agency in biological systems derived from physics (Barp et al., 2022; Friston et al., 2022).

Modern RL algorithms integrate exploratory and exploitative behavior in many different ways. One option is curiosity-driven rewards to encourage exploration. Maximum entropy RL and control-as-inference make decisions by minimizing a KL divergence to the target distribution (Eysenbach & Levine, 2019; Haarnoja et al., 2017, 2018; Levine, 2018; Todorov, 2008; Ziebart et al., 2008), which combines reward maximization with maximum entropy over states. This is similar to active inference on MDPs (Millidge, Tschantz, Seth, et al., 2020). Similarly, the model-free Soft Actor-Critic (Haarnoja et al., 2018) algorithm maximizes both expected reward and entropy. This outperforms other state-of-the-art algorithms in continuous control environments and has been shown to be more sample efficient than its reward-maximizing counterparts (Haarnoja et al., 2018). Hyper (Zintgraf et al., 2021) proposes reward maximization alongside minimizing uncertainty over both external states and model parameters. Bayes-adaptive RL (Guez et al., 2013a, 2013b; Ross et al., 2008, 2011; Zintgraf et al., 2020) provides policies that balance exploration and exploitation with the aim of maximizing reward. Thompson sampling provides a way to balance exploiting current knowledge to maximize immediate performance and accumulating new information to improve future performance (Russo et al., 2017). This reduces to optimizing dual objectives, reward maximization and information gain, similar to active inference on POMDPs. Empirically, Sajid, Ball, et al. (2021) demonstrated that an active inference agent and a Bayesian model-based RL agent using Thompson sampling exhibit similar behavior when preferences are defined over outcomes. They also highlighted that when completely removing the reward signal from the environment, the two agents both select policies that maximize some sort of information gain.

In general, the way each of these approaches the exploration-exploitation dilemma differs in theory and in practice remains largely unexplored.

Appendix B: Proofs

B.1 Proof of Proposition 1. Note that a Bellman optimal state-action policy Π^* is a maximal element according to the partial ordering $\leq$. Existence thus consists of a simple application of Zorn's lemma. Zorn's lemma states that if any increasing chain

$$\Pi_1 \leq \Pi_2 \leq \Pi_3 \leq \dots \tag{B.1}$$

has an upper-bound that is a state-action policy, then there is a maximal element Π^*.

Given the chain equation B.1, we construct an upper bound. We enumerate $\mathbb{A} \times \mathbb{S} \times \mathbb{T}$ by $(\alpha_1, \sigma_1, t_1), \dots, (\alpha_N, \sigma_N, t_N)$. Then the state-action policy sequence

$$\Pi_n(\alpha_1 \mid \sigma_1, t_1), \quad n = 1, 2, 3, \dots$$

is bounded within [0, 1]. By the Bolzano-Weierstrass theorem, there exists a sub-sequence $\Pi_{n_k}(\alpha_1 \mid \sigma_1, t_1)$, $k = 1, 2, 3, \dots$ that converges. Similarly, $\Pi_{n_k}(\alpha_2 \mid \sigma_2, t_2)$ is also a bounded sequence, and by Bolzano-Weierstrass, it has a sub-sequence $\Pi_{n_{k_j}}(a_2 \mid \sigma_2, t_2)$ that converges. We repeatedly take sub-sequences until N. To ease notation, call the resulting sub-sequence Π_m, $m = 1, 2, 3, \dots$

With this, we define $\hat{\Pi} = \lim_{m\to\infty} \Pi_m$. It is straightforward to see that $\hat{\Pi}$ is a state-action policy:

$$\hat{\Pi}(\alpha \mid \sigma, t) = \lim_{m\to\infty} \Pi_m(\alpha \mid \sigma, t) \in [0, 1], \quad \forall(\alpha, \sigma, t) \in \mathbb{A} \times \mathbb{S} \times \mathbb{T},$$

$$\sum_{\alpha\in\mathbb{A}} \hat{\Pi}(\alpha \mid \sigma, t) = \lim_{m\to\infty} \sum_{\alpha\in\mathbb{A}} \Pi_m(\alpha \mid \sigma, t) = 1, \quad \forall(\sigma, t) \in \mathbb{S} \times \mathbb{T}.$$

To show that $\hat{\Pi}$ is an upper bound, take any Π in the original chain of state-action policies, equation B.1. Then by the definition of an increasing sub-sequence, there exists an index $M \in \mathbb{N}$ such that $\forall k \geq M$: $\Pi_k \geq \Pi$. Since limits commute with finite sums, we have $v_{\hat{\Pi}}(s, t) = \lim_{m\to\infty} v_{\Pi_m}(s, t) \geq v_{\Pi_k}(s, t) \geq v_{\Pi}(s, t)$ for any $(s, t) \in \mathbb{S} \times \mathbb{T}$. Thus, by Zorn's lemma, there exists a Bellman, optimal state-action policy Π^*. □

B.2 Proof of Proposition 2. 1) ⇒ 2) : We only need to show assertion *(b)*. By contradiction, suppose that $\exists(s, \alpha) \in \mathbb{S} \times \mathbb{A}$ such that $\Pi(\alpha \mid s, 0) > 0$ and

$$\mathbb{E}_{\Pi}[R(s_{1:T}) \mid s_0 = s, a_0 = \alpha] < \max_{a\in\mathbb{A}} \mathbb{E}_{\Pi}[R(s_{1:T}) \mid s_0 = s, a_0 = a].$$

We let α' be the Bellman optimal action at state s and time 0 defined as

$$\alpha' := \arg\max_{a \in \mathbb{A}} \mathbb{E}_{\Pi}[R(s_{1:T}) \mid s_0 = s, a_0 = a].$$

Then we let Π' be the same state-action policy as Π except that $\Pi'(\cdot \mid s, 0)$ assigns α' deterministically. Then

$$\begin{aligned} v_{\Pi}(s, 0) &= \sum_{a \in \mathbb{A}} \mathbb{E}_{\Pi}[R(s_{1:T}) \mid s_0 = s, a_0 = a]\Pi(a \mid s, 0) \\ &< \max_{a \in \mathbb{A}} \mathbb{E}_{\Pi}[R(s_{1:T}) \mid s_0 = s, a_0 = a] \\ &= \mathbb{E}_{\Pi'}[R(s_{1:T}) \mid s_0 = s, a_0 = \alpha']\Pi'(\alpha' \mid s, 0) \\ &= \sum_{a \in \mathbb{A}} \mathbb{E}_{\Pi'}[R(s_{1:T}) \mid s_0 = s, a_0 = a]\Pi'(a \mid s, 0) \\ &= v_{\Pi'}(s, 0). \end{aligned}$$

So Π is not Bellman optimal, which is a contradiction.

1) $\Leftarrow$ 2) : We only need to show that Π maximizes $v_{\Pi}(s, 0), \forall s \in \mathbb{S}$. By contradiction, there exists a state-action policy Π' and a state $s \in \mathbb{S}$ such that

$$\begin{aligned} & v_{\Pi}(s, 0) < v_{\Pi'}(s, 0), \\ \iff & \sum_{a \in \mathbb{A}} \mathbb{E}_{\Pi}[R(s_{1:T}) \mid s_0 = s, a_0 = a]\Pi(a \mid s, 0) \\ & \quad < \sum_{a \in \mathbb{A}} \mathbb{E}_{\Pi'}[R(s_{1:T}) \mid s_0 = s, a_0 = a]\Pi'(a \mid s, 0). \end{aligned}$$

By a, the left-hand side equals

$$\max_{a \in \mathbb{A}} \mathbb{E}_{\Pi}[R(s_{1:T}) \mid s_0 = s, a_0 = a].$$

Unpacking the expression on the right-hand side,

$$\begin{aligned} & \sum_{a \in \mathbb{A}} \mathbb{E}_{\Pi'}[R(s_{1:T}) \mid s_0 = s, a_0 = a]\Pi'(a \mid s, 0) \\ &= \sum_{a \in \mathbb{A}} \sum_{\sigma \in \mathbb{S}} \mathbb{E}_{\Pi'}[R(s_{1:T}) \mid s_1 = \sigma]P(s_1 = \sigma \mid s_0 = s, a_0 = a)\Pi'(a \mid s, 0) \\ &= \sum_{a \in \mathbb{A}} \sum_{\sigma \in \mathbb{S}} \left\{\mathbb{E}_{\Pi'}[R(s_{2:T}) \mid s_1 = \sigma] + R(\sigma)\right\} P(s_1 = \sigma \mid s_0 = s, a_0 = a)\Pi'(a \mid s, 0) \\ &= \sum_{a \in \mathbb{A}} \sum_{\sigma \in \mathbb{S}} \{v_{\Pi'}(\sigma, 1) + R(\sigma)] P(s_1 = \sigma \mid s_0 = s, a_0 = a)\Pi'(a \mid s, 0). \end{aligned} \tag{B.2}$$

Since Π is Bellman optimal when restricted to $\{1, \ldots, T\}$, we have $v_{\Pi'}(\sigma, 1) \leq v_\Pi(\sigma, 1), \forall \sigma \in \mathbb{S}$. Therefore,

$$\begin{aligned}
&\sum_{a \in \mathbb{A}} \sum_{\sigma \in \mathbb{S}} \{v_{\Pi'}(\sigma, 1) + R(\sigma)] P(s_1 = \sigma \mid s_0 = s, a_0 = a) \Pi'(a \mid s, 0) \\
&\leq \sum_{a \in \mathbb{A}} \sum_{\sigma \in \mathbb{S}} \{v_\Pi(\sigma, 1) + R(\sigma)] P(s_1 = \sigma \mid s_0 = s, a_0 = a) \Pi'(a \mid s, 0).
\end{aligned}$$

Repeating the steps above equation B.2, but in reverse order, yields

$$\begin{aligned}
&\sum_{a \in \mathbb{A}} \mathbb{E}_{\Pi'}[R(s_{1:T}) \mid s_0 = s, a_0 = a] \Pi'(a \mid s, 0) \\
&\quad \leq \sum_{a \in \mathbb{A}} \mathbb{E}_\Pi[R(s_{1:T}) \mid s_0 = s, a_0 = a] \Pi'(a \mid s, 0).
\end{aligned}$$

However,

$$\sum_{a \in \mathbb{A}} \mathbb{E}_\Pi[R(s_{1:T}) \mid s_0 = s, a_0 = a] \Pi'(a \mid s, 0) < \max_{a \in \mathbb{A}} \mathbb{E}_\Pi[R(s_{1:T}) \mid s_0 = s, a_0 = a],$$

which is a contradiction. □

B.3 Proof of Proposition 3. We first prove that state-action policies Π defined as in equation 2.2 are Bellman optimal by induction on T.

$T = 1$:

$$\Pi(a \mid s, 0) > 0 \iff a \in \arg\max_a \mathbb{E}[R(s_1) \mid s_0 = s, a_0 = a], \quad \forall s \in \mathbb{S}$$

is a Bellman optimal state-action policy as it maximizes the total reward possible in the MDP.

Let $T > 1$ be finite and suppose that the proposition holds for MDPs with a temporal horizon of $T - 1$. This means that

$$\begin{aligned}
&\Pi(a \mid s, T-1) > 0 \iff a \in \arg\max_a \mathbb{E}[R(s_T) \mid s_{T-1} = s, a_{T-1} = a], \quad \forall s \in \mathbb{S}, \\
&\Pi(a \mid s, T-2) > 0 \iff a \in \arg\max_a \mathbb{E}_\Pi[R(s_{T-1:T}) \mid s_{T-2} = s, a_{T-2} = a], \\
&\qquad \forall s \in \mathbb{S}, \\
&\qquad \vdots \\
&\Pi(a \mid s, 1) > 0 \iff a \in \arg\max_a \mathbb{E}_\Pi[R(s_{2:T}) \mid s_1 = s, a_1 = a], \quad \forall s \in \mathbb{S},
\end{aligned}$$

is a Bellman optimal state-action policy on the MDP restricted to times 1 to T. Therefore, since

$$\Pi(a \mid s, 0) > 0 \iff a \in \arg\max_{a} \mathbb{E}_{\Pi}[R(s_{1:T}) \mid s_0 = s, a_0 = a], \quad \forall s \in \mathbb{S}.$$

Proposition 2 allows us to deduce that Π is Bellman optimal.

We now show that any Bellman optimal state-action policy satisfies the backward induction algorithm equation 2.2.

Suppose by contradiction that there exists a state-action policy Π that is Bellman optimal but does not satisfy equation 2.2. Say, $\exists (a, s, t) \in \mathbb{A} \times \mathbb{S} \times \mathbb{T}, t < T$, such that

$$\Pi(a \mid s, t) > 0 \text{ and } a \notin \arg\max_{\alpha \in \mathbb{A}} \mathbb{E}_{\Pi}[R(s_{t+1:T}) \mid s_t = s, a_t = \alpha].$$

This implies

$$\mathbb{E}_{\Pi}[R(s_{t+1:T}) \mid s_t = s, a_t = a] < \max_{\alpha \in \mathbb{A}} \mathbb{E}_{\Pi}[R(s_{t+1:T}) \mid s_t = s, a_t = \alpha].$$

Let $\tilde{a} \in \arg\max_{\alpha} \mathbb{E}_{\Pi}[R(s_{t+1:T}) \mid s_t = s, a_t = \alpha]$. Let $\tilde{\Pi}$ be a state-action policy such that $\tilde{\Pi}(\cdot \mid s, t)$ assigns $\tilde{a} \in \mathbb{A}$ deterministically and such that $\tilde{\Pi} = \Pi$ otherwise. Then we can contradict the Bellman optimality of Π as follows:

$$\begin{aligned} v_{\Pi}(s, t) &= E_{\Pi}[R(s_{t+1:T}) \mid s_t = s] \\ &= \sum_{\alpha \in \mathbb{A}} E_{\Pi}[R(s_{t+1:T}) \mid s_t = s, a_t = \alpha] \Pi(\alpha \mid s, t) \\ &< \max_{\alpha \in \mathbb{A}} \mathbb{E}_{\Pi}[R(s_{t+1:T}) \mid s_t = s, a_t = \alpha] \\ &= \mathbb{E}_{\Pi}[R(s_{t+1:T}) \mid s_t = s, a_t = \tilde{a}] \\ &= \mathbb{E}_{\tilde{\Pi}}[R(s_{t+1:T}) \mid s_t = s, a_t = \tilde{a}] \\ &= \sum_{\alpha \in \mathbb{A}} E_{\tilde{\Pi}}[R(s_{t+1:T}) \mid s_t = s, a_t = \alpha] \tilde{\Pi}(\alpha \mid s, t) \\ &= v_{\tilde{\Pi}}(s, t). \end{aligned}$$

□

B.4 Proof of Lemma 1.

$$\begin{aligned} &\lim_{\beta \to +\infty} \arg\min_{\vec{a}} \mathrm{D}_{\mathrm{KL}}[Q(\vec{s} \mid \vec{a}, s_t) \mid C_{\beta}(\vec{s})] \\ &= \lim_{\beta \to +\infty} \arg\min_{\vec{a}} -\mathrm{H}[Q(\vec{s} \mid \vec{a}, s_t)] + \mathbb{E}_{Q(\vec{s} \mid \vec{a}, s_t)}[-\log C_{\beta}(\vec{s})] \\ &= \lim_{\beta \to +\infty} \arg\min_{\vec{a}} -\mathrm{H}[Q(\vec{s} \mid \vec{a}, s_t)] - \beta \mathbb{E}_{Q(\vec{s} \mid \vec{a}, s_t)}[R(\vec{s})] \\ &= \lim_{\beta \to +\infty} \arg\max_{\vec{a}} \mathrm{H}[Q(\vec{s} \mid \vec{a}, s_t)] + \beta \mathbb{E}_{Q(\vec{s} \mid \vec{a}, s_t)}[R(\vec{s})] \end{aligned}$$

$$\begin{aligned}&\subseteq \lim_{\beta\to+\infty} \arg\max_{\vec{a}} \beta\mathbb{E}_{Q(\vec{s}|\vec{a},s_t)}[R(\vec{s})]\\&= \arg\max_{\vec{a}} \mathbb{E}_{Q(\vec{s}|\vec{a},s_t)}[R(\vec{s})].\end{aligned}$$

The inclusion follows from the fact that, as $\beta \to +\infty$, a minimizer of the expected free energy has to maximize $\mathbb{E}_{Q(\vec{s}|\vec{a},s_t)}[R(\vec{s})]$. Among such action sequences, the expected free energy minimizers are those that maximize the entropy of future states $\mathrm{H}[Q(\vec{s} \mid \vec{a}, s_t)]$. □

B.5 Proof of Theorem 1. When $T = 1$, the only action is a_0. We fix an arbitrary initial state $s_0 = s \in \mathbb{S}$. By proposition 2, a Bellman optimal state-action policy is fully characterized by an action a_0^* that maximizes immediate reward:

$$a_0^* \in \arg\max_{a\in\mathbb{A}} \mathbb{E}[R(s_1) \mid s_0 = s, a_0 = a].$$

Recall that by remark 5, this expectation stands for return under the transition probabilities of the MDP:

$$a_0^* \in \arg\max_{a\in\mathbb{A}} \mathbb{E}_{P(s_1|a_0=a,s_0=s)}[R(s_1)].$$

Since transition probabilities are assumed to be known (see equation 3.1), this reads

$$a_0^* \in \arg\max_{a\in\mathbb{A}} \mathbb{E}_{Q(s_1|a_0=a,s_0=s)}[R(s_1)].$$

On the other hand,

$$\begin{aligned}a_0 &\in \lim_{\beta\to+\infty} \arg\max_{a\in\mathbb{A}} \exp(-G(a \mid s_t))\\&= \lim_{\beta\to+\infty} \arg\min_{a\in\mathbb{A}} G(a \mid s_t).\end{aligned}$$

By lemma 1, this implies

$$a_0 \in \arg\max_{a\in\mathbb{A}} \mathbb{E}_{Q(s_1|a_0=a,s_0=s)}[R(s_1)],$$

which concludes the proof. □

B.6 Proof of Theorem 2. We prove this result by induction on the temporal horizon T of the MDP.

The proof of the theorem when $T = 1$ can be seen from the proof of theorem 1. Now suppose that $T > 1$ is finite and that the theorem holds for MDPs with a temporal horizon of $T - 1$.

Our induction hypothesis says that $Q(a_\tau \mid s_\tau)$, as defined in equation 4.8, is a Bellman optimal state-action policy on the MDP restricted to times $\tau = 1, \dots, T$. Therefore, by proposition 2, we only need to show that the action a_0 selected under active inference satisfies

$$a_0 \in \arg\max_{a \in \mathbb{A}} \mathbb{E}_Q[R(\vec{s}) \mid s_0, a_0 = a].$$

This is simple to show as

$$\begin{aligned}
&\arg\max_{a \in \mathbb{A}} \mathbb{E}_Q[R(\vec{s}) \mid s_0, a_0 = a] \\
&= \arg\max_{a \in \mathbb{A}} \mathbb{E}_{P(\vec{s}|a_{1:T}, a_0 = a, s_0) Q(\vec{a}|s_{1:T})}[R(\vec{s})] \quad \text{(by remark 4)} \\
&= \arg\max_{a \in \mathbb{A}} \mathbb{E}_{Q(\vec{s}, \vec{a}|a_0 = a, s_0)}[R(\vec{s})] \quad \text{(as the transitions are known)} \\
&= \lim_{\beta \to +\infty} \arg\max_{a \in \mathbb{A}} \mathbb{E}_{Q(\vec{s}, \vec{a}|a_0 = a, s_0)}[\beta R(\vec{s})] \\
&\supseteq \lim_{\beta \to +\infty} \arg\max_{a \in \mathbb{A}} \mathbb{E}_{Q(\vec{s}, \vec{a}|a_0 = a, s_0)}[\beta R(\vec{s})] - \mathrm{H}[Q(\vec{s} \mid \vec{a}, a_0 = a, s_0)] \\
&= \lim_{\beta \to +\infty} \arg\min_{a \in \mathbb{A}} \mathbb{E}_{Q(\vec{s}, \vec{a}|a_0 = a, s_0)}[-\log C_\beta(\vec{s})] - \mathrm{H}[Q(\vec{s} \mid \vec{a}, a_0 = a, s_0)] \\
&\quad \text{(by equation 4.1)} \\
&= \lim_{\beta \to +\infty} \arg\min_{a \in \mathbb{A}} \mathbb{E}_{Q(\vec{s}, \vec{a}|a_0 = a, s_0)} \mathrm{D_{KL}}[Q(\vec{s} \mid \vec{a}, a_0 = a, s_0) \mid C_\beta(\vec{s})] \\
&= \lim_{\beta \to +\infty} \arg\min_{a \in \mathbb{A}} G(a_0 = a \mid s_0) \quad \text{(by equation 4.6).}
\end{aligned}$$

Therefore, an action a_0 selected under active inference is a Bellman optimal state-action policy on finite temporal horizons. Furthermore, the inclusion follows from the fact that if there are multiple actions that maximize expected reward, that which is selected under active inference maximizes the entropy of beliefs about future states. □

B.7 Proof of Proposition 4. Unpacking the zero temperature limit,

$$\begin{aligned}
&\lim_{\beta \to +\infty} \arg\min_{\vec{a}} G(\vec{a} \mid \tilde{o}) \\
&= \lim_{\beta \to +\infty} \arg\min_{\vec{a}} \mathrm{D_{KL}}[Q(\vec{s} \mid \vec{a}, \tilde{o}) \mid C_\beta(\vec{s})] + \mathbb{E}_{Q(\vec{s}|\vec{a}, \tilde{o})} \mathrm{H}[P(\vec{o} \mid \vec{s})] \\
&= \lim_{\beta \to +\infty} \arg\min_{\vec{a}} -\mathrm{H}[Q(\vec{s} \mid \vec{a}, \tilde{o})] + \mathbb{E}_{Q(\vec{s}|\vec{a}, \tilde{o})}[-\log C_\beta(\vec{s})] + \mathbb{E}_{Q(\vec{s}|\vec{a}, \tilde{o})} \mathrm{H}[P(\vec{o} \mid \vec{s})] \\
&= \lim_{\beta \to +\infty} \arg\min_{\vec{a}} -\mathrm{H}[Q(\vec{s} \mid \vec{a}, \tilde{o})] - \beta \mathbb{E}_{Q(\vec{s}|\vec{a}, \tilde{o})}[R(\vec{s})] + \mathbb{E}_{Q(\vec{s}|\vec{a}, \tilde{o})} \mathrm{H}[P(\vec{o} \mid \vec{s})] \\
&\quad \text{(by equation 4.1)}
\end{aligned}$$

$$\subseteq \lim_{\beta\to+\infty} \arg\max_{\vec{a}} \beta \mathbb{E}_{Q(\vec{s}|\vec{a},\tilde{o})}[R(\vec{s})]$$

$$= \arg\max_{\vec{a}} \mathbb{E}_{Q(\vec{s}|\vec{a},\tilde{o})}[R(\vec{s})].$$

The inclusion follows from the fact that as $\beta \to +\infty$, a minimizer of the expected free energy has first and foremost to maximize $\mathbb{E}_{Q(\vec{s}|\vec{a},\tilde{o})}[R(\vec{s})]$. Among such action sequences, the expected free energy minimizers are those that maximize the entropy of (beliefs about) future states $\mathrm{H}[Q(\vec{s} \mid \vec{a}, \tilde{o})]$ and resolve ambiguity about future outcomes by minimizing $\mathbb{E}_{Q(\vec{s}|\vec{a},\tilde{o})}\mathrm{H}[P(\vec{o} \mid \vec{s})]$.

Acknowledgments

We thank Dimitrije Markovic and Quentin Huys for providing helpful feedback during the preparation of the manuscript.

Funding Information

L.D. is supported by the Fonds National de la Recherche, Luxembourg (Project code: 13568875). N.S. is funded by the Medical Research Council (MR/S502522/1) and 2021-2022 Microsoft PhD Fellowship. K.F. is supported by funding for the Wellcome Centre for Human Neuroimaging (Ref: 205103/Z/16/Z), a Canada-U.K. Artificial Intelligence Initiative (Ref: ES/T01279X/1), and the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreement 945539 (Human Brain Project SGA3). R.S. is supported by the William K. Warren Foundation, the Well-Being for Planet Earth Foundation, the National Institute for General Medical Sciences (P20GM121312), and the National Institute of Mental Health (R01MH123691). This publication is based on work partially supported by the EPSRC Centre for Doctoral Training in Mathematics of Random Systems: Analysis, Modelling and Simulation (EP/S023925/1).

Author Contributions

L.D.: conceptualization, proofs, writing: first draft, review and editing. N.S., T.P., K.F., R.S.: conceptualization, writing: review and editing.

References

Adams, R. A., Stephan, K. E., Brown, H. R., Frith, C. D., & Friston, K. J. (2013). The computational anatomy of psychosis. *Frontiers in Psychiatry*, *4*. 10.3389/fpsyt.2013.00047

Adda, J., & Cooper, R. W. (2003). *Dynamic economics: Quantitative methods and applications*. MIT Press.

Attias, H. (2003). Planning by probabilistic inference. In *Proceedings of the 9th Int. Workshop on Artificial Intelligence and Statistics*.

Barlow, H. B. (1961). *Possible principles underlying the transformations of sensory messages*. MIT Press.

Barlow, H. B. (1974). Inductive inference, coding, perception, and language. *Perception*, *3*(2), 123–134. 10.1068/p030123, PubMed: 4457815

Barp, A., Da Costa, L., França, G., Friston, K., Girolami, M., Jordan, M. I., & Pavliotis, G. A. (2022). Geometric methods for sampling, optimisation, inference and adaptive agents. In F. Nielsen, A. S. R. Srinivasa Rao, & C. Rao (Eds.), *Geometry and statistics*. Elsevier.

Barto, A., Mirolli, M., & Baldassarre, G. (2013). Novelty or surprise? *Frontiers in Psychology*, *4*. 10.3389/fpsyg.2013.00907

Barto, A., & Sutton, R. (1992). *Reinforcement learning: An introduction*. MIT Press.

Beal, M. J. (2003). *Variational algorithms for approximate Bayesian inference*. PhD diss., University of London.

Bellman, R. E. (1957). *Dynamic programming*. Princeton University Press.

Bellman, R. E., & Dreyfus, S. E. (2015). *Applied dynamic programming*. Princeton University Press.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis* (2nd ed.). Springer-Verlag.

Berger-Tal, O., Nathan, J., Meron, E., & Saltz, D. (2014). The exploration-exploitation dilemma: A multidisciplinary framework. *PLOS One*, *9*(4), e95693. 10.1371/journal.pone.0095693

Bertsekas, D. P., & Shreve, S. E. (1996). *Stochastic optimal control: The discrete time case*. Athena Scientific.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, *112*(518), 859–877. 10.1080/01621459.2017.1285773

Botvinick, M., & Toussaint, M. (2012). Planning as inference. *Trends in Cognitive Sciences*, *16*(10), 485–488. 10.1016/j.tics.2012.08.006, PubMed: 22940577

Çatal, O., Nauta, J., Verbelen, T., Simoens, P., & Dhoedt, B. (2019). *Bayesian policy selection using active inference*. http://arxiv.org/abs/1904.08149

Çatal, O., Verbelen, T., Nauta, J., Boom, C. D., & Dhoedt, B. (2020). Learning perception and planning with deep active inference. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 3952–3956). 10.1109/ICASSP40776.2020.9054364

Çatal, O., Verbelen, T., Van de Maele, T., Dhoedt, B., & Safron, A. (2021). Robot navigation as hierarchical active inference. *Neural Networks*, *142*, 192–204. 10.1016/j.neunet.2021.05.010

Champion, T., Bowman, H., & Grzś, M. (2021). *Branching time active inference: Empirical study and complexity class analysis*. http://arxiv.org/abs/2111.11276

Champion, T., Da Costa, L., Bowman, H., & Grześ, M. (2021). *Branching time active inference: The theory and its generality*. http://arxiv.org/abs/2111.11107

Cullen, M., Davey, B., Friston, K. J., & Moran, R. J. (2018). Active inference in OpenAI Gym: A paradigm for computational investigations into psychiatric illness. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *3*(9), 809–818. 10.1016/j.bpsc.2018.06.010, PubMed: 30082215

Da Costa, L., Lanillos, P., Sajid, N., Friston, K., & Khan, S. (2022). How active inference could help revolutionise robotics. *Entropy*, *24*(3), 361. 10.3390/e24030361

Da Costa, L., Parr, T., Sajid, N., Veselic, S., Neacsu, V., & Friston, K. (2020). Active inference on discrete state-spaces: A synthesis. *Journal of Mathematical Psychology*, *99*, 102447. 10.1016/j.jmp.2020.102447

Da Costa, L., Tenka, S., Zhao, D., & Sajid, N. (2022). Active inference as a model of agency. Workshop on RL as a Model of Agency. https://www.sciencedirect.com/science/article/pii/S0022249620300857

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879. 10.1038/nature04766, PubMed: 16778890

Dayan, P., & Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, and Behavioral Neuroscience*, *8*(4), 429–453. 10.3758/CABN.8.4.429

Deci, E., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. Springer.

Eysenbach, B., & Levine, S. (2019). *If MaxEnt Rl is the answer, what is the question?* arXiv:1910.01913.

Fountas, Z., Sajid, N., Mediano, P. A. M. *Deep active inference agents using Monte-Carlo methods*. http://arxiv.org/abs/2006.04176

Friston, K., Da Costa, L., Sajid, N., Heins, C., Ueltzhöffer, K., Pavliotis, G. A., & Parr, T. (2022). *The free energy principle made simpler but not too simple*. http://arxiv.org/abs/2201.06387

Friston, K., Da Costa, L., Hafner, D., Hesp, C., & Parr, T. (2021). Sophisticated inference. *Neural Computation*, *33*(3), 713–763. 10.1162/neco_a_01351, PubMed: 33626312

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., O'Doherty, J., & Pezzulo, G. (2016). Active inference and learning. *Neuroscience and Biobehavioral Reviews*, *68*, 862–879. 10.1016/j.neubiorev.2016.06.022, PubMed: 27375276

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active inference: A process theory *Neural Computation*, *29*(1), 1–49. 10.1162/NECO_a_00912, PubMed: 27870614

Friston, K., Samothrakis, S., & Montague, R. (2012). Active inference and agency: Optimal control without cost functions. *Biological Cybernetics*, *106*(8), 523–541. 10.1007/s00422-012-0512-8, PubMed: 22864468

Friston, K. J., Daunizeau, J., & Kiebel, S. J. (2009). Reinforcement learning or active inference? *PLOS One*, *4*(7), e6421. 10.1371/journal.pone.0006421

Friston, K. J., Daunizeau, J., Kilner, J., & Kiebel, S. J. (2010). Action and behavior: A free-energy formulation. *Biological Cybernetics*, *102*(3), 227–260. 10.1007/s00422-010-0364-z, PubMed: 20148260

Friston, K. J., Lin, M., Frith, C. D., Pezzulo, G., Hobson, J. A., & Ondobaka, S. (2017). Active inference, curiosity and insight. *Neural Computation*, *29*(10), 2633–2683. 10.1162/neco_a_00999, PubMed: 28777724

Friston, K. J., Parr, T., & de Vries, B. (2017). The graphical brain: Belief propagation and active inference. *Network Neuroscience*, *1*(4), 381–414. 10.1162/NETN_a_00018, PubMed: 29417960

Friston, K. J., Rosch, R., Parr, T., Price, C., & Bowman, H. (2018). Deep temporal models and active inference. *Neuroscience and Biobehavioral Reviews*, *90*, 486–501. 10.1016/j.neubiorev.2018.04.004, PubMed: 29747865

Fudenberg, D., & Tirole, J. (1991). *Game theory*. MIT Press.

Gershman, S. J. (2018). Deconstructing the human algorithms for exploration. *Cognition*, *173*, 34–42. 10.1016/j.cognition.2017.12.014, PubMed: 29289795

Gershman, S. J., & Niv, Y. (2010). Learning latent structure: Carving nature at its joints. *Current Opinion in Neurobiology*, *20*(2), 251–256. 10.1016/j.conb.2010.02.008, PubMed: 20227271

Ghavamzadeh, M., Mannor, S., Pineau, J., & Tamar, A. (2016). *Bayesian reinforcement learning: A survey*. arXiv:1609.04436.

Guez, A., Silver, D., & Dayan, P. (2013a). Scalable and efficient Bayes-adaptive reinforcement learning based on Monte-Carlo tree search. *Journal of Artificial Intelligence Research*, *48*, 841–883. 10.1613/jair.4117

Guez, A., Silver, D., & Dayan, P. (2013b). *Efficient Bayes-adaptive reinforcement learning using sample-based search*. http://arxiv.org/abs/1205.3109

Haarnoja, T., Tang, H., Abbeel, P., & Levine, S. (2017). *Reinforcement learning with deep energy-based policies*. arXiv:1702.08165.

Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018). *Soft actor critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor*. CoRR, abs/1801.01290. http://arxiv.org/abs/1801.01290

Huys, Q. J. M., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., & Roiser, J. P. (2012). Bonsai trees in your head: How the Pavlovian system sculpts goal-directed choices by pruning decision trees. *PLOS Computational Biology*, *8*(3), e1002410. 10.1371/journal.pcbi.1002410

Itti, L., & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research*, *49*(10), 1295–1306. 10.1016/j.visres.2008.09.007, PubMed: 18834898

Jaynes, E. T. (1957a). Information theory and statistical mechanics. *Physical Review*, *106*(4), 620–630. 10.1103/PhysRev.106.620

Jaynes, E. T. (1957b). Information theory and statistical mechanics. II. *Physical Review*, *108*(2), 171–190. 10.1103/PhysRev.108.171

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., & Saul, L. K. (1998). An introduction to variational methods for graphical models. In M. I. Jordan (Ed.), *Learning in graphical models* (pp. 105–161). Springer Netherlands. 10.1007/978-94-011-5014-9_5

Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, *101*(1), 99–134. 10.1016/S0004-3702(98)00023-X

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–291. 10.2307/1914185

Kappen, H. J., Gómez, V., & Opper, M. (2012). Optimal control as a graphical model inference problem. *Machine Learning*, *87*(2), 159–182. 10.1007/s10994-012-5278-7

Klyubin, A. S., Polani, D., & Nehaniv, C. L. (2008). Keep your options open: An information-based driving principle for sensorimotor systems. *PLOS One*, *3*(12), e4018. 10.1371/journal.pone.0004018

Lally, N., Huys, Q. J. M., Eshel, N., Faulkner, P., Dayan, P., & Roiser, J. P. (2017). The neural basis of aversive Pavlovian guidance during planning *Journal of Neuroscience*, *37*(42), 10215–10229. 10.1523/JNEUROSCI.0085-17.2017, PubMed: 28924006

Lanillos, P., Pages, J., & Cheng, G. (2020). Robot self/other distinction: Active inference meets neural networks learning in a mirror. In *Proceedings of the European Conference on Artificial Intelligence*.

Levine, S. (2018, May 20). *Reinforcement learning and control as probabilistic inference: tutorial and review*. http://arxiv.org/abs/1805.00909

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *Annals of Mathematical Statistics*, *27*(4), 986–1005. 10.1214/aoms/1177728069

Linsker, R. (1990). Perceptual neural organization: Some approaches based on network models and information theory. *Annual Review of Neuroscience*, *13*(1), 257–281. 10.1146/annurev.ne.13.030190.001353, PubMed: 2183677

MacKay, D. J. C. (2003, September 25). *Information theory, inference and learning algorithms*. Cambridge University Press.

Maisto, D., Gregoretti, F., Friston, K., & Pezzulo, G. (2021, March 25). *Active tree search in large POMDPs*. http://arxiv.org/abs/2103.13860

Marković, D., Stojić, H., Schwöbel, S., & Kiebel, S. J. (2021). An empirical evaluation of active inference in multi-armed bandits. *Neural Networks*, *144*, 229–246. 10.1016/j.neunet.2021.08.018

Mazzaglia, P., Verbelen, T., & Dhoedt, B. (2021). *Contrastive active inference*. https://openreview.net/forum?id=5t5FPwzE6mq

Millidge, B. (2019, March 11). *Implementing predictive processing and active inference: Preliminary steps and results*. PsyArXiv. 10.31234/osf.io/4hb58

Millidge, B. (2020). Deep active inference as variational policy gradients. *Journal of Mathematical Psychology*, *96*, 102348. 10.1016/j.jmp.2020.102348

Millidge, B. (2021). *Applications of the free energy principle to machine learning and neuroscience*. http://arxiv.org/abs/2107.00140

Millidge, B., Tschantz, A., & Buckley, C. L. (2020, April 21). *Whence the expected free energy?* http://arxiv.org/abs/2004.08128

Millidge, B., Tschantz, A., Seth, A. K., & Buckley, C. L. (2020). On the relationship between active inference and control as inference. In T. Verbelen, P. Lanillos, C. L. Buckley, & C. De Boom (Eds.), *Active inference* (pp. 3–11). Springer.

Miranda, M. J., & Fackler, P. L. (2002, September 1). *Applied computational economics and finance*. MIT Press.

Mirza, M. B., Adams, R. A., Mathys, C., & Friston, K. J. (2018). Human visual exploration reduces uncertainty about the sensed world. *PLOS One*, *13*(1), e0190429. 10.1371/journal.pone.0190429

Oliver, G., Lanillos, P., & Cheng, G. (2021). An empirical study of active inference on a humanoid robot. *IEEE Transactions on Cognitive and Developmental Systems PP*(99), 1–1. 10.1109/TCDS.2021.3049907

Optican, L. M., & Richmond, B. J. (1987). Temporal encoding of two-dimensional patterns by single units in primate inferior temporal cortex. III. Information theoretic analysis. *Journal of Neurophysiology*, *57*(1), 162–178. 10.1152/jn.1987.57.1.162, PubMed: 3559670

Oudeyer, P.-Y., & Kaplan, F. (2007). What is intrinsic motivation? A typology of computational approaches. *Frontiers in Neurorobotics*, *1*, 6. 10.3389/neuro.12.006.2007

Parr, T. (2019). *The computational neurology of active vision* (PhD diss.). University College London.

Parr, T., Limanowski, J., Rawji, V., & Friston, K. (2021). The computational neurology of movement under active inference. *Brain*, *144*(6), 1799–1818. 10.1093/brain/awab085, PubMed: 33704439

Parr, T., Markovic, D., Kiebel, S. J., & Friston, K. J. (2019). Neuronal message passing using mean-field, Bethe, and marginal approximations. *Scientific Reports*, *9*(1), 1889. 10.1038/s41598-018-38246-3

Parr, T., Pezzulo, G., & Friston, K. J. (2022, March 29). *Active inference: The free energy principle in mind, brain, and behavior*. MIT Press.

Paul, A., Sajid, N., Gopalkrishnan, M., & Razi, A. (2021, August 27). *Active inference for stochastic control*. http://arxiv.org/abs/2108.12245

Pavliotis, G. A. (2014). *Stochastic processes and applications: Diffusion processes, the Fokker-Planck and Langevin equations*. Springer.

Pearl, J. (1998). Graphical models for probabilistic and causal reasoning. In P. Smets (Ed.), *Quantified representation of uncertainty and imprecision* (pp. 367–389). Springer Netherlands.

Pezzato, C., Ferrari, R., & Corbato, C. H. (2020). A novel adaptive controller for robot manipulators based on active inference. *IEEE Robotics and Automation Letters*, *5*(2), 2973–2980. 10.1109/LRA.2020.2974451

Pio-Lopez, L., Nizard, A., Friston, K., & Pezzulo, G. (2016). Active inference and robot control: A case study. *Journal of the Royal Society Interface*, *13*(122), 20160616. 10.1098/rsif.2016.0616

Puterman, M. L. (2014, August 28). *Markov decision processes: Discrete stochastic dynamic programming*. Wiley.

Rahme, J., & Adams, R. P. (2019, June 24). *A theoretical connection between statistical physics and reinforcement learning*. http://arxiv.org/abs/1906.10228

Rawlik, K., Toussaint, M., & Vijayakumar, S. (2013). On stochastic optimal control and reinforcement learning by approximate inference. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*. https://www.aaai.org/ocs/index.php/IJCAI/IJCAI13/paper/view/6658

Ross, S., Chaib-draa, B., & Pineau, J. (2008). Bayes-adaptive POMDPs. In J. C. Platt, D. Koller, Y. Singer, & S. T. Roweis (Eds.), *Advances in neural information processing systems*, *20* (pp. 1225–1232). Curran. http://papers.nips.cc/paper/3333-bayes-adaptive-pomdps.pdf

Ross, S., Pineau, J., Chaib-draa, B., & Kreitmann, P. (2011). A Bayesian approach for learning and planning in partially observable Markov decision processes. *Journal of Machine Learning Research*, *12* (2011).

Russo, D., & Van Roy, B. (2014). Learning to optimize via posterior sampling. *Mathematics of Operations Research*, *39*(4), 1729–1770. 10.1287/moor.2014.0650

Russo, D., & Van Roy, B. (2016). An information-theoretic analysis of Thompson sampling. *Journal of Machine Learning Research*, *17*(1), 2442–2471.

Russo, D., Van Roy, B., Kazerouni, A., Osband, I., & Wen, Z. (2017). *A tutorial on Thompson sampling*. arXiv:1707.02038.

Sajid, N., Ball, P. J., Parr, T., & Friston, K. J. (2021). Active inference: Demystified and compared. *Neural Computation*, *33*(3), 674–712. 10.1162/neco_a_01357, PubMed: 33400903

Sajid, N., Holmes, E., Costa, L. D., Price, C., & Friston, K. (2022). *A mixed generative model of auditory word repetition*. 10.1101/2022.01.20.477138

Sajid, N., Tigas, P., Zakharov, A., Fountas, Z., & Friston, K. (2021, July 18). *Exploration and preference satisfaction trade-off in reward-free learning*. http://arxiv.org/abs/2106.04316

Sales, A. C., Friston, K. J., Jones, M. W., Pickering, A. E., & Moran, R. J. (2019). Locus Coeruleus tracking of prediction errors optimises cognitive flexibility: An active inference model. *PLOS Computational Biology*, *15*(1), e1006267. 10.1371/journal.pcbi.1006267

Sancaktar, C., van Gerven, M., & Lanillos, P. (2020, May 29). *End-to-end pixel-based deep active inference for body perception and action*. http://arxiv.org/abs/2001.05847

Sargent, R. W. H. (2000). Optimal control. *Journal of Computational and Applied Mathematics*, *124*(1), 361–371. 10.1016/S0377-0427(00)00418-0

Schmidhuber, J. (2006). Developmental robotics, optimal artificial curiosity, creativity, music, and the fine arts. *Connection Science*, *18*(2), 173–187. 10.1080/09540090600768658

Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, *2*(3), 230–247. 10.1109/TAMD.2010.2056368

Schneider, T., Belousov, B., Abdulsamad, H., & Peters, J. (2022, June 1). *Active inference for robotic manipulation*. 10.48550/arXiv.2206.10313

Schulz, E., & Gershman, S. J. (2019). The algorithmic architecture of exploration in the human brain. *Current Opinion in Neurobiology*, *55*, 7–14. 10.1016/j.conb.2018.11.003, PubMed: 30529148

Schwartenbeck, P., FitzGerald, T. H. B., Mathys, C., Dolan, R., & Friston, K. (2015). The dopaminergic midbrain encodes the expected certainty about desired outcomes. *Cerebral Cortex*, *25*(10), 3434–3445. 10.1093/cercor/bhu159, PubMed: 25056572

Schwartenbeck, P., FitzGerald, T. H. B., Mathys, C., Dolan, R., Kronbichler, M., & Friston, K. (2015). Evidence for surprise minimization over value maximization in choice behavior. *Scientific Reports*. *5*, 16575. 10.1038/srep16575

Schwartenbeck, P., Passecker, J., Hauser, T. U., FitzGerald, T. H., Kronbichler, M., & Friston, K. J. (2019). Computational mechanisms of curiosity and goal-directed exploration. *eLife*, *45*.

Shoham, Y., Powers, R., & Grenager, T. (2003). *Multi-agent reinforcement learning: A critical survey*. Computer Science Department, Stanford University.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., . . . Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489. 10.1038/nature16961, PubMed: 26819042

Smith, R., Friston, K. J., & Whyte, C. J. (2022). A step-by-step tutorial on active inference and its application to empirical data. *Journal of Mathematical Psychology*, *107*, 102632. 10.1016/j.jmp.2021.102632

Smith, R., Khalsa, S. S., & Paulus, M. P. (2021). An active inference approach to dissecting reasons for nonadherence to antidepressants. *Biological Psychiatry.*

Cognitive Neuroscience and Neuroimaging, *6*(9), 919–934. 10.1016/j.bpsc.2019.11.012, PubMed: 32044234

Smith, R., Kirlic, N., Stewart, J. L., Touthang, J., Kuplicki, R., Khalsa, S. S. F., . . . Aupperle, R. L. (2021). Greater decision uncertainty characterizes a transdiagnostic patient sample during approach-avoidance conflict: A computational modeling approach. *Journal of Psychiatry an Neuroscience*, *46*(1), E74–E87. 10.1503/jpn.200032

Smith, R., Kirlic, N., Stewart, J. L., Touthang, J., Kuplicki, R., McDermott, T. J., . . . Aupperle, R. L. (2021). Long-term stability of computational parameters during approach-avoidance conflict in a transdiagnostic psychiatric patient sample. *Scientific Reports*, *11*(1), 11783. 10.1038/s41598-021-91308-x

Smith, R., Kuplicki, R., Feinstein, J., Forthman, K. L., Stewart, J. L., Paulus, M. P., . . . Khalsa, S. S. (2020). A Bayesian computational model reveals a failure to adapt interoceptive precision estimates across depression, anxiety, eating, and substance use disorders. *PLOS Computational Biology*, *16*(12), e1008484. 10.1371/journal.pcbi.1008484

Smith, R., Kuplicki, R., Teed, A., Upshaw, V., & Khalsa, S. S. (2020, September 29). *Confirmatory evidence that healthy individuals can adaptively adjust prior expectations and interoceptive precision estimates.* 10.1101/2020.08.31.275594

Smith, R., Mayeli, A., Taylor, S., Al Zoubi, O., Naegele, J., & Khalsa, S. S. (2021). Gut inference: A computational modeling approach. *Biological Psychology, 164* 108152. 10.1016/j.biopsycho.2021.108152

Smith, R., Schwartenbeck, P., Parr, T., & Friston, K. J. (2019). *An active inference model of concept learning*. bioRxiv:633677. 10.1101/633677

Smith, R., Schwartenbeck, P., Parr, T., & Friston, K. J. (2020). An active inference approach to modeling structure learning: concept learning as an example case. *Frontiers in Computational Neuroscience, 14*. 10.3389/fncom.2020.00041

Smith, R., Schwartenbeck, P., Stewart, J. L., Kuplicki, R., Ekhtiari, H., & Paulus, M. P. (2020). Imprecise action selection in substance use disorder: evidence for active learning impairments when solving the explore-exploit dilemma. *Drug and Alcohol Dependence*, *215*, 108208. 10.1016/j.drugalcdep.2020.108208

Smith, R., Taylor, S., Stewart, J. L., Guinjoan, S. M., Ironside, M., Kirlic, N., . . . Paulus, M. P. (2022). Slower learning rates from negative outcomes in substance use disorder over a 1-year period and their potential predictive utility. *Computational Psychiatry*, *6*(1), 117–141. 10.5334/cpsy.85

Still, S., & Precup, D. (2012). An information-theoretic approach to curiosity-driven reinforcement learning. *Theory in Biosciences*, *131*(3), 139–148. 10.1007/s12064-011-0142-z, PubMed: 22791268

Stolle, M., & Precup, D. (2002). Learning options in reinforcement learning. *Lecture Notes in Computer Science*, 212–223. Springer. 10.1007/3-540-45622-8_16

Stone, J. V. (2015, February 1). *Information theory: A tutorial introduction*. Sebtel Press.

Stone, J. V. (2019). *Artificial intelligence engines: A tutorial introduction to the mathematics of deep learning*. Sebtel Press.

Sun, Y., Gomez, F., & Schmidhuber, J. (2011, March 29). *Planning to be surprised: Optimal Bayesian exploration in dynamic environments*. http://arxiv.org/abs/1103.5708

Tanaka, T. (1999). A theory of mean field approximation. In S. Solla, T. Leen, & K. Müller (Eds.), *Advances in neural information processing systems*, *11*. MIT Press.

Tervo, D. G. R., Tenenbaum, J. B., & Gershman, S. J. (2016). Toward the neural implementation of structure learning. *Current Opinion in Neurobiology*, *37*, 99–105. 10.1016/j.conb.2016.01.014, PubMed: 26874471

Todorov, E. (2006). Linearly-solvable Markov decision problems. In *Advances in neural information processing systems, 19*. MIT Press. https://papers.nips.cc/paper/2006/hash/d806ca13ca3449af72a1ea5aedbed26a-Abstract.html

Todorov, E. (2008). General duality between optimal control and estimation. In *Proceedings of the 47th IEEE Conference on Decision and Control* (pp. 4286–4292). 10.1109/CDC.2008.4739438

Todorov, E. (2009). Efficient computation of optimal actions. In *Proceedings of the National Academy of Sciences*, *106*(28), 11478–11483. 10.1073/pnas.0710743106

Tokic, M., & Palm, G. (2011). Value-difference based exploration: Adaptive Control between epsilon-greedy and Softmax. In J. Bach & S. Edelkamp (Eds.), *KI 2011: Advances in artificial intelligence* (pp. 335–346). Springer. 10.1007/978-3-642-24455-1_33

Toussaint, M. (2009). Robot trajectory optimization using approximate inference. In *Proceedings of the 26th Annual International Conference on Machine Learning* (pp. 1049–1056). 10.1145/1553374.1553508

Tschantz, A., Baltieri, M., Seth, A. K., & Buckley, C. L. (2019, November 24). *Scaling active inference*. http://arxiv.org/abs/1911.10601

Tschantz, A., Millidge, B., Seth, A. K., & Buckley, C. L. (2020). *Reinforcement learning through active inference*. http://arxiv.org/abs/2002.12636

Tschantz, A., Seth, A. K., & Buckley, C. L. (2020). Learning action-oriented models through active inference. *PLOS Computational Biology*, *16*(4), e1007805. 10.1371/journal.pcbi.1007805

van den Broek, B., Wiegerinck, W., & Kappen, B. (2010). *Risk sensitive path integral control*. https://arxiv.org/ftp/arxiv/papers/1203/1203.3523.pdf

van der Himst, O., & Lanillos, P. (2020). *Deep active inference for partially observable MDPs*. 10.1007/978-3-030-64919-7_8

Von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton University Press.

Wainwright, M. J., & Jordan, M. I. (2007). Graphical models, exponential families, and variational inference. *Foundations and Trend in Machine Learning*, *1*(1–2), 1–305. 10.1561/2200000001

Wilson, R. C., Bonawitz, E., Costa, V. D., & Ebitz, R. B. (2021). Balancing exploration and exploitation with information and randomization. *Current Opinion in Behavioral Sciences*, *38*, 49–56. 10.1016/j.cobeha.2020.10.001, PubMed: 33184605

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology. General*, *143*(6), 2074–2081. 10.1037/a0038199

Xu, H. A., Modirshanechi, A., Lehmann, M. P., Gerstner, W., & Herzog, M. H. (2021). Novelty is not surprise: Human exploratory and adaptive behavior in sequential decision-making. *PLOS Computational Biology*, *17*(6), e1009070. 10.1371/journal.pcbi.1009070

Zermelo, E. (1913). *Über eine Anwendung der Mengenlehre auf die Theorie des Schachspiels*. https://www.mathematik.uni-muenchen.de/∼spielth/artikel/Zermelo.pdf

Ziebart, B. (2010). *Modeling purposeful adaptive behavior with the principle of maximum causal entropy.* Carnegie Mellon University.

Ziebart, B. D., Maas, A. L., Bagnell, J. A., & Dey, A. K. (2008). Maximum entropy inverse reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence.*

Zintgraf, L., Shiarlis, K., Igl, M., Schulze, S., Gal, Y., Hofmann, K., & Whiteson, S. (2020, February 27). *VariBAD: A very good method for Bayes-adaptive deep RL via meta- learning*. http://arxiv.org/abs/1910.08348

Zintgraf, L. M., Feng, L., Lu, C., Igl, M., Hartikainen, K., Hofmann, K., & Whiteson, S. (2021). Exploration in approximate hyper-state space for meta reinforcement learning. *International Conference on Machine Learning* (pp. 12991–13001).

Received September 13, 2022; accepted December 17, 2022.